Web Appendix: TM-OKC: An Unsupervised Topic Model for Text in Online Knowledge Communities

Web Appendix A: Summary of Topic Modeling Applications in OKC Research

Table A1 summarizes major studies about the topic modeling applications in OKC. To do so, we conduct an extensive literature search among premier journals in IS and other business disciplines, and reviewed recent studies that apply topic models on texts with Q&A relations and threaded structures.

Table A1: Summary of Topic Modeling Applications on OKC Texts

Ways to use topic vectors	Authors (Year)	Topic model applied	Textual data	Research topic
As independent variables	Yue et al. (2019)	LDA	Threaded posts in an online hacking forum	Extracting topics discussed in online hacking forum posts and exploring the impact of topics on distributed denial of service attacks
	Narang et al. (2022)	LDA	Online learning discussions	Identifying the impact of content type (i.e., topic) on learner engagement
	Gour et al. (2022)	LDA	Social media discussions	Deriving topics from social media discussions to help predict the disease outbreak
As dependent variables	Singh et al. (2014)	LDA	Enterprise blogs and comments	Impact of various factors (e.g., textual characteristics) on users' blog-reading of different topics
	Li et al. (2016)	LDA	Threaded advertisements in a cyber-carding community	Profiling key sellers using the derived topic vectors of advertisement threads
	Geva et al. (2019)	LDA	Threaded Tweets	Identifying users' interested topics from their blogs and studying the user behavior of shaping online persona via retweets
As control variables	Bapna et al. (2019)	LDA	Posts and comments in online brand communities	Impact of posts content dimensions on user engagement with topics as control variables
	Xie et al. (2020)	LDA	Online Bitcoin-related discussion threads	Controlling the topics of discussion threads to identify the role of network cohesion in predicting Bitcoin returns
	Kumar et al. (2022)	LDA	Posts and comments in online brand communities	Impact of trademarking hashtags on social media consumer engagement with topics as control variables
Deriving new variables	Lappas et al. (2016)	LDA	Customers' reviews and businesses' responses	Using the derived topics to further classify businesses' responses to customers' comments into different types

	Guo et al. (2017)	LDA, hLDA, and DTM	Articles and comments on a blogging platform	Using the derived topics to extract representative information
	Samtani et al. (2017)	LDA	Posts and comments in an online hacker forum	Building topic-specific social networks based on the topics learned from texts
	Hwang et al. (2019)	LDA	Q&A posts in a customer support crowdsourcing community	Constructing each user's information network based on learned topics to further explore how topic-based information network influences the generation of novel ideas.
	Kokkodis et al. (2020)	LDA	Q&A posts in an online diabetes community	Categorizing users into different contribution types by the topics of their posts
	Mousavi et al. (2020)	HDP	Q&A posts in a health-related community	Calculating the topic similarity between an answer and the question based on the derived topic vectors
	Pu et al. (2020)	LDA	Q&A posts on an enterprise platform	Impact of identity disclosure on users' effort measured by the topic similarity of Q&A
	Bachura et al. (2022)	LDA	Threaded Tweets	Extracting breach-related concepts based on top salient terms from each topic
	Kyriakou et al. (2022)	CTM	Descriptions and comments of product designs in an online innovation community	Using derived topic vectors to calculate the similarity among product designs to further measure novelty
	Oh et al. (2022)	LDA	News articles	Combining learned topics with sentiment analysis to calculate topic valence
	Pu et al. (2022)	LDA	Q&A posts on an enterprise platform	Effects of hierarchy on question answering, controlling user's knowledge level measured by the topic similarity between the question and the user's existing answers

References

Bachura, E., Valecha, R., Chen, R., and Rao, H. R. 2022. "The OPM Data Breach: An Investigation of Shared Emotional Reactions On Twitter," *MIS Quarterly* (46:2), pp. 881–910.

Bapna, S., Benner, M.J., and Qiu, L. 2019. "Nurturing Online Communities: An Empirical Investigation," *MIS Quarterly* (43:2), pp. 425–452.

Geva, H., Oestreicher-Singer, G., and Saar-Tsechansky, M. 2019. "Using Retweets When Shaping Our Online Persona: Topic Modeling Approach," *MIS Quarterly* (43:2), pp. 501–524.

Gour, A., Aggarwal, S., and Kumar, S. 2022. "Lending Ears to Unheard Voices: An Empirical Analysis of User-Generated Content on Social Media," *Production and Operations Management* (31:6), pp. 2457–2476.

Guo, X., Wei, Q., Chen, G., Zhang, J., and Qiao, D. 2017. "Extracting Representative Information on Intra-Organizational Blogging Platforms," *MIS Quarterly* (41:4), pp. 1105–1128.

Hwang, E.H., Singh, P.V., and Argote, L. 2019. "Jack of All, Master of Some: Information Network and Innovation in Crowdsourcing Communities," *Information Systems Research* (30:2), pp. 389–410.

Kokkodis, M., Lappas, T., and Ransbotham, S. 2020. "From Lurkers to Workers: Predicting Voluntary Contribution and Community Welfare," *Information Systems Research* (31:2), pp. 607–626.

Kumar, N., Qiu, L., and Kumar, S. 2022. "A Hashtag is Worth a Thousand Words: An Empirical Investigation of Social Media Strategies in Trademarking Hashtags," *Information Systems Research* (33:4), pp. 1403–1427.

Kyriakou, H., Nickerson, J. V., and Majchrzak, A. 2022. "Novelty and the Structure of Design Landscapes: A Relational View of Online Innovation Communities," *MIS Quarterly* (46:3), pp. 1691–1720.

Lappas, T., Sabnis, G., and Valkanas, G. 2016. "The Impact of Fake Reviews on Online Visibility: A Vulnerability Assessment of the Hotel Industry," *Information Systems Research* (27:4), pp. 940–961.

Li, W., Chen, H., and Nunamaker Jr, J. F. 2016. "Identifying and Profiling Key Sellers in Cyber Carding Community: Azsecure Text Mining System," *Journal of Management Information Systems* (33:4), pp. 1059–1086.

Mousavi, R., Raghu, T. S., and Frey, K. 2020. "Harnessing Artificial Intelligence to Improve the Quality of Answers in Online Question-Answering Health Forums," *Journal of Management Information Systems* (37:4), pp. 1073–1098.

Narang, U., Yadav, M. S., and Rindfleisch, A. 2022. "The "Idea Advantage": How Content Sharing Strategies Impact Engagement in Online Learning Platforms," *Journal of Marketing Research* (59:1), pp. 61–78.

Oh, H., Goh, K. Y., and Phan, T. Q. 2022. "Are You What You Tweet? The Impact of Sentiment on Digital News Consumption and Social Media Sharing," *Information Systems Research*.

Pu, J., Chen, Y., Qiu, L., and Cheng, H. K. 2020. "Does Identity Disclosure Help or Hurt User Content Generation? Social Presence, Inhibition, and Displacement Effects," *Information Systems Research* (31:2), pp. 297–322.

Pu, J., Liu, Y., Chen, Y., Qiu, L., and Cheng, H.K. 2022. "What Questions Are You Inclined to Answer? Effects of Hierarchy in Corporate Q&A Communities," *Information Systems Research* (33:1), pp. 244–264.

Samtani, S., Chinn, R., Chen, H., and Nunamaker Jr, J. F. 2017. "Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence," *Journal of Management Information Systems* (34:4), pp. 1023–1053.

Singh, P.V., Sahoo, N., and Mukhopadhyay, T. 2014. "How to Attract and Retain Readers in Enterprise Blogging?" *Information Systems Research* (25:1), pp. 35–52.

Xie, P., Chen, H., and Hu, Y. J. 2020. "Signal or Noise in Social Media Discussions: The Role of Network Cohesion in Predicting the Bitcoin Market," *Journal of Management Information Systems* (37:4), pp. 933–956.

Yue, W.T., Wang, Q., and Hui, K.-L. 2019. "See No Evil, Hear No Evil? Dissecting the Impact of Online Hacker Forums," *MIS Quarterly* (43:1), pp. 73–95.

Web Appendix B: Details of Variational Inference

B.1. The Objective Function: ELBO

The ELBO is given in Equation (3) of Section 3.3. Before deriving optimization procedures in the **coordinate ascent algorithm**, we write each term of the ELBO in a specific functional form as follows.

(1) The first term (1):

$$E_u\left[\log p\left(\boldsymbol{\eta_q}; \boldsymbol{\mu}, \boldsymbol{\Sigma_q}\right)\right] = \frac{1}{2}\log\left|\boldsymbol{\Sigma_q^{-1}}\right| - \frac{K}{2}\log 2\pi - \frac{1}{2}E_u\left[\left(\boldsymbol{\eta_q}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma_q^{-1}}\left(\boldsymbol{\eta_q}-\boldsymbol{\mu}\right)\right],$$

where

$$E_u\left[(\boldsymbol{\eta_q}-\boldsymbol{\mu})^T\boldsymbol{\Sigma_q^{-1}}(\boldsymbol{\eta_q}-\boldsymbol{\mu})\right] = Tr\left[diag(\boldsymbol{\sigma_q})^2\boldsymbol{\Sigma_q^{-1}}\right] + \left(\boldsymbol{\lambda_q}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma_q^{-1}}\left(\boldsymbol{\lambda_q}-\boldsymbol{\mu}\right).$$

(2) The second term (N_q):

$$E_u\left[\log p\left(z_q^{n_q}|\boldsymbol{\eta_q}\right)\right] = E_u\left[\boldsymbol{\eta_q^T}\boldsymbol{\phi_q^{n_q}}\right] - E_u\left[\log\left(\sum_{k=1}^{K} e^{\boldsymbol{\eta_q^k}}\right)\right] = \sum_{k=1}^{K}\lambda_q^k\phi_q^{n_q,k} - E_u\left[\log\left(\sum_{k=1}^{K} e^{\boldsymbol{\eta_q^k}}\right)\right].$$

As a logistic normal distribution is not conjugate to multinomial distribution, this term $E_u\left[\log\left(\sum_{k=1}^{K} e^{\boldsymbol{\eta_q^k}}\right)\right]$ cannot be analytically computed. To preserve the lower bound on the log probability, we use a Taylor expansion here:

$$E_u\left[\log\left(\sum_{k=1}^{K} e^{\boldsymbol{\eta_q^k}}\right)\right] \le \xi_q^{-1}\left(\sum_{k=1}^{K} E_u\left(e^{\boldsymbol{\eta_q^k}}\right)\right) + \log\xi_q - 1.$$

As $E_u\left(e^{\boldsymbol{\eta_q^k}}\right) = e^{\boldsymbol{\lambda_q^k}+\frac{1}{2}\left(\sigma_q^k\right)^2}$, we know that:

$$E_u\left[\log p\left(z_q^{n_q}|\boldsymbol{\eta_q}\right)\right] \ge \sum_{k=1}^{K}\lambda_q^k\phi_q^{n_q,k} - \xi_q^{-1}\left(\sum_{k=1}^{K} e^{\lambda_q^k+\frac{1}{2}\left(\sigma_q^k\right)^2}\right) - \log\xi_q + 1.$$

(3) The third term (N_q):

$$E_u\left[\log p\left(w_q^{n_q} \middle| z_q^{n_q}, \boldsymbol{\beta}_{\boldsymbol{q}}^{\boldsymbol{1:K}}\right)\right] = \sum_{k=1}^{K} \phi_q^{n_q,k} E_u\left(\log \beta_q^{\mathrm{k},n_q}\right) = \sum_{k=1}^{K} \phi_q^{n_q,k}\left[\psi\left(\tau_q^{k,n_q}\right) - \psi\left(\sum_{v=1}^{V} \tau_q^{k,v}\right)\right].$$

(4) The fourth term (K):

$E_u\left[\log p\left(\boldsymbol{\beta}_{\boldsymbol{q}}^{\boldsymbol{k}}; \boldsymbol{\alpha}_{\boldsymbol{q}}\right)\right] = -\log B\left(\boldsymbol{\alpha}_{\boldsymbol{q}}\right) + \sum_{v=1}^{V}(\alpha_q^v - 1)\left[\psi\left(\tau_q^{k,v}\right) - \psi\left(\sum_{v=1}^{V} \tau_q^{k,v}\right)\right],$

where V is the vocabulary size.

(5) The fifth term (1):

$$E_u[\log p(\boldsymbol{x}_{\boldsymbol{d}}; \boldsymbol{\delta})] = -\log B(\boldsymbol{\delta}) + \sum_{i=1}^{2}(\delta_i - 1)\left[\psi\left(\nu_d^i\right) - \psi\left(\sum_{i=1}^{2} \nu_d^i\right)\right].$$

(6) The sixth term (T):

$$E_u\left[\log p\left(\boldsymbol{y}_{\boldsymbol{a}_{\boldsymbol{t}}} \middle| \boldsymbol{x}_{\boldsymbol{d}}\right)\right] = \sum_{i=1}^{2} \psi_{a_t}^i E_u\left[\log\left(x_d^i\right)\right] = \sum_{i=1}^{2} \psi_{a_t}^i\left[\psi\left(\nu_d^i\right) - \psi\left(\sum_{i=1}^{2} \nu_d^i\right)\right].$$

(7) The seventh term (T):

$E_u\left[\log p\left(\boldsymbol{\eta}_{\boldsymbol{a}_{\boldsymbol{t}}} \middle| \boldsymbol{\eta}_{\boldsymbol{q}}, \overline{\boldsymbol{\eta}}_{\boldsymbol{a}_{\boldsymbol{t-1}}}, \boldsymbol{y}_{\boldsymbol{a}_{\boldsymbol{t}}}; \boldsymbol{\Sigma}_{\boldsymbol{a}_{\boldsymbol{f}}}, \boldsymbol{\Sigma}_{\boldsymbol{a}_{\boldsymbol{n}}}, \gamma\right)\right]$

$$= \psi_{a_t}^1\left\{\frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\boldsymbol{a}_{\boldsymbol{f}}}^{\boldsymbol{-1}}\right| - \frac{K}{2}\log 2\pi\right.$$

$$\left. - \frac{1}{2}E_u\left[\left(\boldsymbol{\eta}_{\boldsymbol{a}_{\boldsymbol{t}}} - \frac{\boldsymbol{\eta}_{\boldsymbol{q}} + \gamma\overline{\boldsymbol{\eta}}_{\boldsymbol{a}_{\boldsymbol{t-1}}}}{1+\gamma}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{a}_{\boldsymbol{f}}}^{\boldsymbol{-1}}\left(\boldsymbol{\eta}_{\boldsymbol{a}_{\boldsymbol{t}}} - \frac{\boldsymbol{\eta}_{\boldsymbol{q}} + \gamma\overline{\boldsymbol{\eta}}_{\boldsymbol{a}_{\boldsymbol{t-1}}}}{1+\gamma}\right)\right]\right\}$$

$$+ \psi_{a_t}^2\left\{\frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\boldsymbol{a}_{\boldsymbol{n}}}^{\boldsymbol{-1}}\right| - \frac{K}{2}\log 2\pi - \frac{1}{2}E_u\left[\left(\boldsymbol{\eta}_{\boldsymbol{a}_{\boldsymbol{t}}} - \boldsymbol{\eta}_{\boldsymbol{q}}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{a}_{\boldsymbol{n}}}^{\boldsymbol{-1}}\left(\boldsymbol{\eta}_{\boldsymbol{a}_{\boldsymbol{t}}} - \boldsymbol{\eta}_{\boldsymbol{q}}\right)\right]\right\}.$$

Let $\mathrm{A} = E_u\left[\left(\boldsymbol{\eta}_{\boldsymbol{a}_{\boldsymbol{t}}} - \frac{\boldsymbol{\eta}_{\boldsymbol{q}} + \gamma\overline{\boldsymbol{\eta}}_{\boldsymbol{a}_{\boldsymbol{t-1}}}}{1+\gamma}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{a}_{\boldsymbol{f}}}^{\boldsymbol{-1}}\left(\boldsymbol{\eta}_{\boldsymbol{a}_{\boldsymbol{t}}} - \frac{\boldsymbol{\eta}_{\boldsymbol{q}} + \gamma\overline{\boldsymbol{\eta}}_{\boldsymbol{a}_{\boldsymbol{t-1}}}}{1+\gamma}\right)\right]$, then:

If $\mathrm{t} = 1$,

$A = Tr\left[diag(\boldsymbol{\sigma}_{\boldsymbol{a}_{\boldsymbol{t}}})^2 \boldsymbol{\Sigma}_{\boldsymbol{a}_{\boldsymbol{f}}}^{\boldsymbol{-1}}\right] + Tr\left[diag(\boldsymbol{\sigma}_{\boldsymbol{q}})^2 \boldsymbol{\Sigma}_{\boldsymbol{a}_{\boldsymbol{f}}}^{\boldsymbol{-1}}\right] + \left(\boldsymbol{\lambda}_{\boldsymbol{a}_{\boldsymbol{t}}} - \boldsymbol{\lambda}_{\boldsymbol{q}}\right)^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{a}_{\boldsymbol{f}}}^{\boldsymbol{-1}}\left(\boldsymbol{\lambda}_{\boldsymbol{a}_{\boldsymbol{t}}} - \boldsymbol{\lambda}_{\boldsymbol{q}}\right).$

If $\mathrm{t} \geq 2$,

$$A = Tr\left[diag(\boldsymbol{\sigma}_{\boldsymbol{a}_t})^2\boldsymbol{\Sigma}_{\boldsymbol{a}_f}^{-\mathbf{1}}\right] + \frac{1}{(1+\gamma)^2}Tr\left[diag(\boldsymbol{\sigma}_{\boldsymbol{q}})^2\boldsymbol{\Sigma}_{\boldsymbol{a}_f}^{-\mathbf{1}}\right] + \sum_{i=1}^{t-1}\left(\frac{\gamma\zeta_i^t}{1+\gamma}\right)^2 Tr\left[diag(\boldsymbol{\sigma}_{\boldsymbol{a}_i})^2\boldsymbol{\Sigma}_{\boldsymbol{a}_f}^{-\mathbf{1}}\right]$$

$$+\left(\boldsymbol{\lambda}_{\boldsymbol{a}_t} - \frac{1}{1+\gamma}\boldsymbol{\lambda}_{\boldsymbol{q}} - \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{t}-\mathbf{1}}\frac{\gamma\zeta_i^t}{1+\gamma}\boldsymbol{\lambda}_{\boldsymbol{a}_i}\right)^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{a}_f}^{-\mathbf{1}}\left(\boldsymbol{\lambda}_{\boldsymbol{a}_t} - \frac{1}{1+\gamma}\boldsymbol{\lambda}_{\boldsymbol{q}} - \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{t}-\mathbf{1}}\frac{\gamma\zeta_i^t}{1+\gamma}\boldsymbol{\lambda}_{\boldsymbol{a}_i}\right).$$

Let B $= E_u\left[\left(\boldsymbol{\eta}_{\boldsymbol{a}_t} - \boldsymbol{\eta}_{\boldsymbol{q}}\right)^T\boldsymbol{\Sigma}_{\boldsymbol{a}_n}^{-\mathbf{1}}\left(\boldsymbol{\eta}_{\boldsymbol{a}_t} - \boldsymbol{\eta}_{\boldsymbol{q}}\right)\right]$, then:

$$B = Tr\left[diag(\boldsymbol{\sigma}_{\boldsymbol{a}_t})^2\boldsymbol{\Sigma}_{\boldsymbol{a}_n}^{-\mathbf{1}}\right] + Tr\left[diag(\boldsymbol{\sigma}_{\boldsymbol{q}})^2\boldsymbol{\Sigma}_{\boldsymbol{a}_n}^{-\mathbf{1}}\right] + \left(\boldsymbol{\lambda}_{\boldsymbol{a}_t} - \boldsymbol{\lambda}_{\boldsymbol{q}}\right)^{\boldsymbol{T}}\boldsymbol{\Sigma}_{\boldsymbol{a}_n}^{-\mathbf{1}}\left(\boldsymbol{\lambda}_{\boldsymbol{a}_t} - \boldsymbol{\lambda}_{\boldsymbol{q}}\right).$$

(8) The eighth term ($\sum_{t=1}^{T} N_{a_t}$, similar to the second term):

$$E_u\left[\log p\left(z_{a_t}^{n_{a_t}}\middle|\boldsymbol{\eta}_{\boldsymbol{a}_t}\right)\right] = E_u\left[\boldsymbol{\eta}_{\boldsymbol{a}_t}^{\boldsymbol{T}}\boldsymbol{\phi}_{\boldsymbol{a}_t}^{\boldsymbol{n}_{\boldsymbol{a}_t}}\right] - E_u\left[\log\left(\sum_{k=1}^{K} e^{\eta_{a_t}^k}\right)\right]$$

$$= \sum_{k=1}^{K}\lambda_{a_t}^k\phi_{a_t}^{n_{a_t},k} - E_u\left[\log\left(\sum_{k=1}^{K} e^{\boldsymbol{\eta}_{a_t}^{\boldsymbol{k}}}\right)\right]$$

$$\geq \sum_{k=1}^{K}\lambda_{a_t}^k\phi_{a_t}^{n_{a_t},k} - \xi_{a_t}^{-1}\left(\sum_{k=1}^{K} e^{\lambda_{a_t}^k+\frac{1}{2}\left(\sigma_{a_t}^k\right)^2}\right) - \log\xi_{a_t} + 1.$$

(9) The ninth term ($\sum_{t=1}^{T} N_{a_t}$, similar to the third term):

$$E_u\left[\log p\left(w_{a_t}^{n_{a_t}}\middle|z_{a_t}^{n_{a_t}}, \boldsymbol{\beta}_{\boldsymbol{a}}^{\mathbf{1}:\boldsymbol{K}}\right)\right] = \sum_{k=1}^{K}\phi_{a_t}^{n_{a_t},k}\, E_u\left(\log\beta_a^{\mathrm{k},n_{a_t}}\right)$$

$$= \sum_{k=1}^{K}\phi_{a_t}^{n_{a_t},k}\left[\psi\left(\tau_a^{k,n_{a_t}}\right) - \psi\left(\sum_{v=1}^{V}\tau_a^{k,v}\right)\right].$$

(10) The tenth term (K, similar to the fourth term):

$$E_u\left[\log p\left(\boldsymbol{\beta}_{\boldsymbol{a}}^{\boldsymbol{k}}; \boldsymbol{\alpha}_{\boldsymbol{a}}\right)\right] = -\log B(\boldsymbol{\alpha}_{\boldsymbol{a}}) + \sum_{v=1}^{V}(\alpha_a^v - 1)\left[\psi\left(\tau_a^{k,v}\right) - \psi\left(\sum_{v=1}^{V}\tau_a^{k,v}\right)\right].$$

(11) The eleventh term (1, the entropy term):

$$
\begin{aligned}
H(u) = {} & \sum_{k=1}^{K} \frac{1}{2} \left[\log\left(\sigma_q^k\right)^2 + \log 2\pi + 1 \right] + \sum_{t=1}^{T} \sum_{k=1}^{K} \frac{1}{2} \left[\log\left(\sigma_{a_t}^k\right)^2 + \log 2\pi + 1 \right] + \log B(\boldsymbol{\nu_d}) \\
& - \sum_{i=1}^{2} \left(\nu_d^i - 1\right) \left[\psi\left(\nu_d^i\right) - \psi\left(\sum_{i=1}^{2} \nu_d^i \right) \right] - \sum_{t=1}^{T} \sum_{i=1}^{2} \psi_{a_t}^i \log \psi_{a_t}^i \\
& - \sum_{n_q=1}^{N_q} \sum_{k=1}^{K} \phi_q^{n_q,k} \log \phi_q^{n_q,k} - \sum_{t=1}^{T} \sum_{n_{a_t}=1}^{N_{a_t}} \sum_{k=1}^{K} \phi_{a_t}^{n_{a_t},k} \log \phi_{a_t}^{n_{a_t},k} + \sum_{k=1}^{K} \log B\left(\boldsymbol{\tau_q^k}\right) \\
& - \sum_{k=1}^{K} \sum_{v=1}^{V} (\tau_q^{k,v} - 1) \left[\psi\left(\tau_q^{k,v}\right) - \psi\left(\sum_{v=1}^{V} \tau_q^{k,v} \right) \right] + \sum_{k=1}^{K} \log B\left(\boldsymbol{\tau_a^k}\right) \\
& - \sum_{k=1}^{K} \sum_{v=1}^{V} (\tau_a^{k,v} - 1) \left[\psi\left(\tau_a^{k,v}\right) - \psi\left(\sum_{v=1}^{V} \tau_q^{k,v} \right) \right].
\end{aligned}
$$

To summarize, the ELBO is given by:

$$ELBO = \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-\mathbf{1}}\right| - \frac{K}{2}\log 2\pi - \frac{1}{2}Tr\left[diag\left(\boldsymbol{\sigma}_{\boldsymbol{q}}\right)^{2}\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-\mathbf{1}}\right] - \frac{1}{2}\left(\boldsymbol{\lambda}_{\boldsymbol{q}} - \boldsymbol{\mu}\right)^{T}\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-\mathbf{1}}\left(\boldsymbol{\lambda}_{\boldsymbol{q}} - \boldsymbol{\mu}\right)$$

$$+ \sum_{n_q=1}^{N_q}\left\{\sum_{k=1}^{K}\lambda_q^k\phi_q^{n_q,k} - \xi_q^{-1}\left(\sum_{k=1}^{K}e^{\lambda_q^k+\frac{1}{2}\left(\sigma_q^k\right)^2}\right) - \log\xi_q + 1\right\}$$

$$+ \sum_{n_q=1}^{N_q}\left\{\sum_{k=1}^{K}\phi_q^{n_q,k}\left[\psi\left(\tau_q^{k,n_q}\right) - \psi\left(\sum_{v=1}^{V}\tau_q^{k,v}\right)\right]\right\}$$

$$+ \sum_{k=1}^{K}\left\{-\log B\left(\boldsymbol{\alpha}_{\boldsymbol{q}}\right) + \sum_{v=1}^{V}\left(\alpha_q^v - 1\right)\left[\psi\left(\tau_q^{k,v}\right) - \psi\left(\sum_{v=1}^{V}\tau_q^{k,v}\right)\right]\right\} - \log B(\boldsymbol{\delta})$$

$$+ \sum_{i=1}^{2}\left(\delta_i - 1\right)\left[\psi\left(\nu_d^i\right) - \psi\left(\sum_{i=1}^{2}\nu_d^i\right)\right] + \sum_{t=1}^{T}\sum_{i=1}^{2}\psi_{a_t}^i\left[\psi\left(\nu_d^i\right) - \psi\left(\sum_{i=1}^{2}\nu_d^i\right)\right]$$

$$+ \sum_{t=1}^{T}\left\{\psi_{a_t}^1\left\{\frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\boldsymbol{a_f}}^{-\mathbf{1}}\right| - \frac{K}{2}\log 2\pi\right.\right.$$

$$\left.- \frac{1}{2}E_u\left[\left(\boldsymbol{\eta}_{\boldsymbol{a_t}} - \frac{\boldsymbol{\eta}_{\boldsymbol{q}} + \gamma\overline{\boldsymbol{\eta}}_{\boldsymbol{a_{t-1}}}}{1+\gamma}\right)^{T}\boldsymbol{\Sigma}_{\boldsymbol{a_f}}^{-\mathbf{1}}\left(\boldsymbol{\eta}_{\boldsymbol{a_t}} - \frac{\boldsymbol{\eta}_{\boldsymbol{q}} + \gamma\overline{\boldsymbol{\eta}}_{\boldsymbol{a_{t-1}}}}{1+\gamma}\right)\right]\right\}$$

$$\left.+ \psi_{a_t}^2\left\{\frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\boldsymbol{a_n}}^{-\mathbf{1}}\right| - \frac{K}{2}\log 2\pi - \frac{1}{2}E_u\left[\left(\boldsymbol{\eta}_{\boldsymbol{a_t}} - \boldsymbol{\eta}_{\boldsymbol{q}}\right)^{T}\boldsymbol{\Sigma}_{\boldsymbol{a_n}}^{-\mathbf{1}}\left(\boldsymbol{\eta}_{\boldsymbol{a_t}} - \boldsymbol{\eta}_{\boldsymbol{q}}\right)\right]\right\}\right\}$$

$$+ \sum_{t=1}^{T}\sum_{n_{a_t}=1}^{N_{a_t}}\left\{\sum_{k=1}^{K}\lambda_{a_t}^k\phi_{a_t}^{n_{a_t},k} - \xi_{a_t}^{-1}\left(\sum_{k=1}^{K}e^{\lambda_{a_t}^k+\frac{1}{2}\left(\sigma_{a_t}^k\right)^2}\right) - \log\xi_{a_t} + 1\right\}$$

$$+ \sum_{t=1}^{T}\sum_{n_{a_t}=1}^{N_{a_t}}\left\{\sum_{k=1}^{K}\phi_{a_t}^{n_{a_t},k}\left[\psi\left(\tau_a^{k,n_{a_t}}\right) - \psi\left(\sum_{v=1}^{V}\tau_a^{k,v}\right)\right]\right\}$$

$$+ \sum_{k=1}^{K}\left\{-\log B\left(\boldsymbol{\alpha}_{\boldsymbol{a}}\right) + \sum_{v=1}^{V}\left(\alpha_a^v - 1\right)\left[\psi\left(\tau_a^{k,v}\right) - \psi\left(\sum_{v=1}^{V}\tau_a^{k,v}\right)\right]\right\}$$

$$+\sum_{k=1}^{K}\frac{1}{2}\left[\log\left(\sigma_q^k\right)^2+\log 2\pi+1\right]+\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{1}{2}\left[\log\left(\sigma_{a_t}^k\right)^2+\log 2\pi+1\right]$$

$$+\log B(\boldsymbol{\nu_d})-\sum_{i=1}^{2}\left(\nu_d^i-1\right)\left[\psi\left(\nu_d^i\right)-\psi\left(\sum_{i=1}^{2}\nu_d^i\right)\right]-\sum_{t=1}^{T}\sum_{i=1}^{2}\psi_{a_t}^i\log\psi_{a_t}^i$$

$$-\sum_{n_q=1}^{N_q}\sum_{k=1}^{K}\phi_q^{n_q,k}\log\phi_q^{n_q,k}-\sum_{t=1}^{T}\sum_{n_{a_t}=1}^{N_{a_t}}\sum_{k=1}^{K}\phi_{a_t}^{n_{a_t},k}\log\phi_{a_t}^{n_{a_t},k}+\sum_{k=1}^{K}\log B\left(\boldsymbol{\tau_q^k}\right)$$

$$-\sum_{k=1}^{K}\sum_{v=1}^{V}\left(\tau_q^{k,v}-1\right)\left[\psi\left(\tau_q^{k,v}\right)-\psi\left(\sum_{v=1}^{V}\tau_q^{k,v}\right)\right]+\sum_{k=1}^{K}\log B\left(\boldsymbol{\tau_a^k}\right)$$

$$-\sum_{k=1}^{K}\sum_{v=1}^{V}\left(\tau_a^{k,v}-1\right)\left[\psi\left(\tau_a^{k,v}\right)-\psi\left(\sum_{v=1}^{V}\tau_q^{k,v}\right)\right].$$

B.2. Optimization Algorithm

For optimization in the variational inference, we use the **coordinate ascent algorithm**, iteratively maximizing the ELBO with respect to each variational parameter. Here we derive to update these parameters, including $\boldsymbol{\xi_q}, \boldsymbol{\phi_q}, \boldsymbol{\lambda_q}, \boldsymbol{\sigma_q}, \boldsymbol{\xi_{a_t}}, \boldsymbol{\phi_{a_t}}, \boldsymbol{\lambda_{a_t}}, \boldsymbol{\sigma_{a_t}}, \boldsymbol{\psi_{a_t}}, \boldsymbol{\nu_d}, \boldsymbol{\tau_q}, \boldsymbol{\tau_a}$.

(1) Maximize with respect to ξ_q

$$\frac{dELBO}{d\xi_q}=N_q\left[\xi_q^{-2}\left(\sum_{k=1}^{K}e^{\lambda_q^k+\frac{1}{2}\left(\sigma_q^k\right)^2}\right)-\xi_q^{-1}\right].$$

Set $\frac{dELBO}{d\xi_q}=0$, we obtain: $\xi_q^*=\sum_{k=1}^{K}e^{\lambda_q^k+\frac{1}{2}\left(\sigma_q^k\right)^2}$. Note that this is $E_q\left(\sum_{k=1}^{K}e^{\boldsymbol{\eta_q^k}}\right)$.

(2) Maximize with respect to $\boldsymbol{\phi_q^{n_q}}$

$$\frac{dELBO}{d\phi_q^{n_q,k}}=\lambda_q^k+\left[\psi\left(\tau_q^{k,n_q}\right)-\psi\left(\sum_{v=1}^{V}\tau_q^{k,v}\right)\right]-\log\phi_q^{n_q,k}-1.$$

Since $\sum_{k=1}^{K} \phi_q^{n_q,k} = 1$, $\frac{dELBO}{d\phi_q^{n_q,k}} = 0$ implies $\phi_q^{n_q,k} \propto e^{\lambda_q^k + \left[\psi\left(\tau_q^{k,n_q}\right) - \psi\left(\sum_{v=1}^{V} \tau_q^{k,v}\right)\right]}$, so then we can write the updating formula as follows:

$$\phi_q^{n_q,k} = \frac{e^{\lambda_q^k + \left[\psi\left(\tau_q^{k,n_q}\right) - \psi\left(\sum_{v=1}^{V} \tau_q^{k,v}\right)\right]}}{\sum_{k=1}^{K} e^{\lambda_q^k + \left[\psi\left(\tau_q^{k,n_q}\right) - \psi\left(\sum_{v=1}^{V} \tau_q^{k,v}\right)\right]}}.$$

(3) Maximize with respect to $\boldsymbol{\lambda_q}$

$$\frac{dELBO}{d\boldsymbol{\lambda_q}} = -\boldsymbol{\Sigma_q^{-1}}(\boldsymbol{\lambda_q} - \boldsymbol{\mu}) + \sum_{n_q=1}^{N_q} \boldsymbol{\phi_q^{n_q}} - \frac{N_q}{\xi_q}\left(e^{\boldsymbol{\lambda_q} + \frac{1}{2}(\boldsymbol{\sigma_q})^2}\right) + \psi_{a_1}^1 \boldsymbol{\Sigma_{a_f}^{-1}}(\boldsymbol{\lambda_{a_1}} - \boldsymbol{\lambda_q})$$

$$+ \frac{1}{1+\gamma} \sum_{t=2}^{T} \psi_{a_t}^1 \boldsymbol{\Sigma_{a_f}^{-1}} \left(\boldsymbol{\lambda_{a_t}} - \frac{1}{1+\gamma}\boldsymbol{\lambda_q} - \sum_{i=1}^{t-1} \frac{\gamma \zeta_i^t}{1+\gamma} \boldsymbol{\lambda_{a_i}}\right) + \sum_{t=1}^{T} \psi_{a_t}^2 \boldsymbol{\Sigma_{a_n}^{-1}}(\boldsymbol{\lambda_{a_t}} - \boldsymbol{\lambda_q}).$$

Set $\frac{dELBO}{d\boldsymbol{\lambda_q}} = 0$, and we find that this cannot be analytically solved. In previous studies, numerical methods have often been applied in variational inference for nonconjugate models (Blei and Lafferty 2005; Blei and Lafferty 2007; Wang and Blei 2013; Roberts et al. 2016). Thus, we use the extended limited memory BFGS (Byrd et al. 1995; Wang et al. 2021) algorithm to update $\boldsymbol{\lambda_q}$ numerically. Moreover, note that as the Hessian matrix of ELBO on $\boldsymbol{\lambda_q}$ is **negative-definite**, this is a convex optimization problem, which guarantees that we can achieve the global optimum with the numerical method:

$$H(ELBO)_{\boldsymbol{\lambda_q}} = -\boldsymbol{\Sigma_q^{-1}} - \frac{N_q}{\xi_q}\left(e^{\boldsymbol{\lambda_q} + \frac{1}{2}(\boldsymbol{\sigma_q})^2}\right) - \psi_{a_1}^1 \boldsymbol{\Sigma_{a_f}^{-1}} - \frac{1}{(1+\gamma)^2} \sum_{t=2}^{T} \psi_{a_t}^1 \boldsymbol{\Sigma_{a_f}^{-1}} - \sum_{t=1}^{T} \psi_{a_t}^2 \boldsymbol{\Sigma_{a_n}^{-1}}.$$

(4) Maximize with respect to $\boldsymbol{\sigma_q}$

$$\frac{dELBO}{d\left(\sigma_q^k\right)^2} = -\frac{1}{2}\boldsymbol{\Sigma}_{\boldsymbol{q}}^{\boldsymbol{-1}(\boldsymbol{k},\boldsymbol{k})} - \frac{N_q}{2\xi_q}\left(e^{\lambda_{\mathrm{q}}^k + \frac{1}{2}\left(\sigma_{\mathrm{q}}^k\right)^2}\right) - \frac{1}{2}\psi_{a_1}^1 \boldsymbol{\Sigma}_{\boldsymbol{a_f}}^{\boldsymbol{-1}(\boldsymbol{k},\boldsymbol{k})} - \frac{1}{2}\sum_{\boldsymbol{t=2}}^{\boldsymbol{T}} \frac{\psi_{a_t}^1}{(1+\gamma)^2} \boldsymbol{\Sigma}_{\boldsymbol{a_f}}^{\boldsymbol{-1}(\boldsymbol{k},\boldsymbol{k})}$$

$$-\frac{1}{2}\sum_{\boldsymbol{t=1}}^{\boldsymbol{T}} \psi_{a_t}^2 \boldsymbol{\Sigma}_{\boldsymbol{a_n}}^{\boldsymbol{-1}(\boldsymbol{k},\boldsymbol{k})} + \frac{1}{2\left(\sigma_q^k\right)^2}.$$

Similar to $\boldsymbol{\lambda_q}$, we use the extended limited memory BFGS to update $\boldsymbol{\sigma_q}$. This is also a convex optimization problem since the second derivative of ELBO on $\left(\sigma_q^k\right)^2$ is negative:

$$\frac{d^2 ELBO}{\left[d\left(\sigma_q^k\right)^2\right]^2} = -\frac{N_q}{4\xi_q}\left(e^{\lambda_{\mathrm{q}}^k + \frac{1}{2}\left(\sigma_{\mathrm{q}}^k\right)^2}\right) - \frac{1}{2\left[\left(\sigma_q^k\right)^2\right]^2}.$$

(5) Maximize with respect to ξ_{a_t}, $\mathrm{t} = 1{:}\,\mathrm{T}$

$$\frac{dELBO}{d\xi_{a_t}} = N_a\left[\xi_{a_t}^{-2}\left(\sum_{k=1}^{K} e^{\lambda_{a_t}^k + \frac{1}{2}\left(\sigma_{a_t}^k\right)^2}\right) - \xi_{a_t}^{-1}\right].$$

Set $\frac{dELBO}{d\xi_{a_t}} = 0$, we obtain $\xi_{a_t}^* = \sum_{k=1}^{K} e^{\lambda_{a_t}^k + \frac{1}{2}\left(\sigma_{a_t}^k\right)^2}$. Note that this is $E_q\left(\sum_{k=1}^{K} e^{\boldsymbol{\eta}_{a_t}^{\boldsymbol{k}}}\right)$.

(6) Maximize with respect to $\boldsymbol{\phi}_{\boldsymbol{a_t}}^{\boldsymbol{n_{a_t}}}$, $\mathrm{t} = 1{:}\,\mathrm{T}$

$$\frac{dELBO}{d\phi_{a_t}^{n_{a_t},k}} = \lambda_{a_t}^k + \left[\psi\left(\tau_a^{k,n_{a_t}}\right) - \psi\left(\sum_{v=1}^{V} \tau_a^{k,v}\right)\right] - \log \phi_{\boldsymbol{a_t}}^{n_{a_t},k} - 1.$$

Since $\sum_{k=1}^{K} \phi_{a_t}^{n_{a_t},k} = 1$,

$$\phi_{a_t}^{n_{a_t},k} \propto e^{\lambda_{a_t}^k + \left[\psi\left(\tau_a^{k,n_{a_t}}\right) - \psi\left(\sum_{v=1}^{V} \tau_a^{k,v}\right)\right]}, \text{specifically } \phi_{a_t}^{n_{a_t},k} = \frac{e^{\lambda_{a_t}^k + \left[\psi\left(\tau_a^{k,n_{a_t}}\right) - \psi\left(\sum_{v=1}^{V} \tau_a^{k,v}\right)\right]}}{\sum_{k=1}^{K} e^{\lambda_{a_t}^k + \left[\psi\left(\tau_a^{k,n_{a_t}}\right) - \psi\left(\sum_{v=1}^{V} \tau_a^{k,v}\right)\right]}}.$$

(7) Maximize with respect to $\boldsymbol{\lambda}_{a_t}$, $\mathrm{t} = 1{:}\,\mathrm{T}$

When $\mathrm{t} = 1$:

$$\frac{dELBO}{d\boldsymbol{\lambda}_{\boldsymbol{a_t}}} = \psi_{a_t}^{1}\left[-\boldsymbol{\Sigma}_{\boldsymbol{a_f}}^{-\boldsymbol{1}}\left(\boldsymbol{\lambda}_{\boldsymbol{a_t}} - \boldsymbol{\lambda}_{\boldsymbol{q}}\right) + \sum_{j=t+1}^{T}\frac{\gamma\zeta_t^j}{1+\gamma}\boldsymbol{\Sigma}_{\boldsymbol{a_f}}^{-\boldsymbol{1}}\left(\boldsymbol{\lambda}_{\boldsymbol{a_j}} - \frac{1}{1+\gamma}\boldsymbol{\lambda}_{\boldsymbol{q}} - \sum_{i=1}^{j-1}\frac{\gamma\zeta_i^j}{1+\gamma}\boldsymbol{\lambda}_{\boldsymbol{a_i}}\right)\right]$$

$$+ \psi_{a_t}^{2}\left[-\boldsymbol{\Sigma}_{\boldsymbol{a_n}}^{-\boldsymbol{1}}\left(\boldsymbol{\lambda}_{\boldsymbol{a_t}} - \boldsymbol{\lambda}_{\boldsymbol{q}}\right)\right] + \sum_{n_{a_t}=1}^{N_{a_t}}\boldsymbol{\phi}_{a_t}^{\boldsymbol{n}_{a_t}} - \frac{N_{a_t}}{\xi_{a_t}}\left(e^{\boldsymbol{\lambda}_{\boldsymbol{a_t}}+\frac{1}{2}\left(\boldsymbol{\sigma}_{\boldsymbol{a_t}}\right)^2}\right).$$

When t ≥ 2:

$$\frac{dELBO}{d\boldsymbol{\lambda}_{a_t}} = \psi_{a_t}^{1}\left[-\boldsymbol{\Sigma}_{\boldsymbol{a_f}}^{-\boldsymbol{1}}\left(\boldsymbol{\lambda}_{\boldsymbol{a_t}} - \frac{1}{1+\gamma}\boldsymbol{\lambda}_{\boldsymbol{q}} - \sum_{i=1}^{t-1}\frac{\gamma\zeta_i^t}{1+\gamma}\boldsymbol{\lambda}_{\boldsymbol{a_i}}\right)\right.$$

$$\left. + \sum_{j=t+1}^{T}\frac{\gamma\zeta_t^j}{1+\gamma}\boldsymbol{\Sigma}_{\boldsymbol{a_f}}^{-\boldsymbol{1}}\left(\boldsymbol{\lambda}_{\boldsymbol{a_j}} - \frac{1}{1+\gamma}\boldsymbol{\lambda}_{\boldsymbol{q}} - \sum_{i=1}^{j-1}\frac{\gamma\zeta_i^j}{1+\gamma}\boldsymbol{\lambda}_{\boldsymbol{a_i}}\right)\right] + \psi_{a_t}^{2}\left[-\boldsymbol{\Sigma}_{\boldsymbol{a_n}}^{-\boldsymbol{1}}\left(\boldsymbol{\lambda}_{\boldsymbol{a_t}} - \boldsymbol{\lambda}_{\boldsymbol{q}}\right)\right]$$

$$+ \sum_{n_{a_t}=1}^{N_{a_t}}\boldsymbol{\phi}_{a_t}^{\boldsymbol{n}_{a_t}} - \frac{N_{a_t}}{\xi_{a_t}}\left(e^{\boldsymbol{\lambda}_{\boldsymbol{a_t}}+\frac{1}{2}\left(\boldsymbol{\sigma}_{\boldsymbol{a_t}}\right)^2}\right).$$

Similar to $\boldsymbol{\lambda}_{\boldsymbol{q}}$, this is a convex optimization problem because the Hessian matrix of ELBO on $\boldsymbol{\lambda}_{a_t}$ is **negative-definite**:

$$H(ELBO)_{\lambda_{a_t}} = -\psi_{a_t}^{1}\boldsymbol{\Sigma}_{\boldsymbol{a_f}}^{-\boldsymbol{1}} - \psi_{a_t}^{1}\sum_{j=t+1}^{T}\left(\frac{\gamma\zeta_t^j}{1+\gamma}\right)^2\boldsymbol{\Sigma}_{\boldsymbol{a_f}}^{-\boldsymbol{1}} - \psi_{a_t}^{2}\boldsymbol{\Sigma}_{\boldsymbol{a_n}}^{-\boldsymbol{1}} - \frac{N_{a_t}}{\xi_{a_t}}\left(e^{\boldsymbol{\lambda}_{\boldsymbol{a_t}}+\frac{1}{2}\left(\boldsymbol{\sigma}_{\boldsymbol{a_t}}\right)^2}\right).$$

Therefore, we use the extended limited memory BFGS to update $\boldsymbol{\lambda}_{a_t}$. From this updating process, we can see that the topic distribution of a focal answer is impacted by not only the focal question and former answers, but also the latter answers. In other words, our Bayesian framework takes the bi-directional correlations of threaded answers into account, which enhances the capability of posterior model inference.

(8) Maximize with respect to $\boldsymbol{\sigma}_{\boldsymbol{a_t}}$, $\mathrm{t} = 1{:}\,\mathrm{T}$

$$\frac{dELBO}{d\left(\sigma_{a_t}^k\right)^2} = \psi_{a_t}^1\left[-\frac{1}{2}\boldsymbol{\Sigma}_{\boldsymbol{a_f}}^{\boldsymbol{-1}(\boldsymbol{k},\boldsymbol{k})} - \frac{1}{2}\sum_{j=t+1}^{T}\left(\frac{\gamma\zeta_t^j}{1+\gamma}\right)^2 \boldsymbol{\Sigma}_{\boldsymbol{a_f}}^{\boldsymbol{-1}(\boldsymbol{k},\boldsymbol{k})}\right] + \psi_{a_t}^2\left[-\frac{1}{2}\boldsymbol{\Sigma}_{\boldsymbol{a_n}}^{\boldsymbol{-1}(\boldsymbol{k},\boldsymbol{k})}\right]$$

$$-\frac{N_{a_t}}{2\xi_{a_t}}\left(e^{\lambda_{a_t}^k+\frac{1}{2}\left(\sigma_{a_t}^k\right)^2}\right) + \frac{1}{2\left(\sigma_{a_t}^k\right)^2}.$$

Similar to $\boldsymbol{\sigma_q}$, this is a convex optimization problem because the second derivative of ELBO on $\left(\sigma_{a_t}^k\right)^2$ is negative:

$$\frac{d^2ELBO}{\left[d\left(\sigma_{a_t}^k\right)^2\right]^2} = -\frac{N_{a_t}}{4\xi_{a_t}}\left(e^{\lambda_{a_t}^k+\frac{1}{2}\left(\sigma_{a_t}^k\right)^2}\right) - \frac{1}{2\left[\left(\sigma_{a_t}^k\right)^2\right]^2},$$

so we use the extended limited memory BFGS to update $\boldsymbol{\sigma_{a_t}}$.

(9) Maximize with respect to $\boldsymbol{\psi_{a_t}}$, $\mathrm{t} = 1{:}\,\mathrm{T}$

$$\frac{dELBO}{d\psi_{a_t}^1} = \left[\psi(\nu_d^1) - \psi\left(\sum_{i=1}^{2}\nu_d^i\right)\right] + \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\boldsymbol{a_f}}^{\boldsymbol{-1}}\right| - \frac{K}{2}\log 2\pi$$

$$-\frac{1}{2}E_u\left[\left(\boldsymbol{\eta_{a_t}} - \frac{\boldsymbol{\eta_q} + \gamma\overline{\boldsymbol{\eta}}_{\boldsymbol{a_{t-1}}}}{1+\gamma}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{a_f}}^{\boldsymbol{-1}}\left(\boldsymbol{\eta_{a_t}} - \frac{\boldsymbol{\eta_q} + \gamma\overline{\boldsymbol{\eta}}_{\boldsymbol{a_{t-1}}}}{1+\gamma}\right)\right] - \log\psi_{a_t}^1 - 1.$$

$$\frac{dELBO}{d\psi_{a_t}^2} = \left[\psi(\nu_d^2) - \psi\left(\sum_{i=1}^{2}\nu_d^i\right)\right] + \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\boldsymbol{a_n}}^{\boldsymbol{-1}}\right| - \frac{K}{2}\log 2\pi - \frac{1}{2}E_u\left[\left(\boldsymbol{\eta_{a_t}} - \boldsymbol{\eta_q}\right)^T\boldsymbol{\Sigma}_{\boldsymbol{a_n}}^{\boldsymbol{-1}}\left(\boldsymbol{\eta_{a_t}} - \boldsymbol{\eta_q}\right)\right]$$

$$-\log\psi_{a_t}^2 - 1.$$

Since $\sum_{i=1}^{2}\psi_{a_t}^i = 1$, it is easy to know $\psi_{a_t}^1 = \frac{e^C}{e^C+e^D}$, $\psi_{a_t}^2 = \frac{e^D}{e^C+e^D}$, where $\mathrm{C} = \left[\psi(\nu_d^1) - \psi\left(\sum_{i=1}^{2}\nu_d^i\right)\right] + \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\boldsymbol{a_f}}^{\boldsymbol{-1}}\right| - \frac{1}{2}E_u\left[\left(\boldsymbol{\eta_{a_t}} - \frac{\boldsymbol{\eta_q}+\gamma\overline{\boldsymbol{\eta}}_{\boldsymbol{a_{t-1}}}}{1+\gamma}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{a_f}}^{\boldsymbol{-1}}\left(\boldsymbol{\eta_{a_t}} - \frac{\boldsymbol{\eta_q}+\gamma\overline{\boldsymbol{\eta}}_{\boldsymbol{a_{t-1}}}}{1+\gamma}\right)\right]$, $D = \left[\psi(\nu_d^2) - \psi\left(\sum_{i=1}^{2}\nu_d^i\right)\right] + \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\boldsymbol{a_n}}^{\boldsymbol{-1}}\right| - \frac{1}{2}E_u\left[\left(\boldsymbol{\eta_{a_t}} - \boldsymbol{\eta_q}\right)^T\boldsymbol{\Sigma}_{\boldsymbol{a_n}}^{\boldsymbol{-1}}\left(\boldsymbol{\eta_{a_t}} - \boldsymbol{\eta_q}\right)\right]$.

(10) Maximize with respect to $\boldsymbol{\nu_d}$

$$\frac{dELBO}{d\nu_d^i} = (\delta_i - 1)\left[\psi_1\left(\nu_d^i\right) - \psi_1\left(\sum_{i=1}^{2} \nu_d^i\right)\right] + \sum_{t=1}^{T} \psi_{a_t}^i \left[\psi_1\left(\nu_d^i\right) - \psi_1\left(\sum_{i=1}^{2} \nu_d^i\right)\right]$$

$$+ \left[\psi\left(\nu_d^i\right) - \psi\left(\sum_{i=1}^{2} \nu_d^i\right)\right] - \left[\psi\left(\nu_d^i\right) - \psi\left(\sum_{i=1}^{2} \nu_d^i\right)\right]$$

$$- \left(\nu_d^i - 1\right)\left[\psi_1\left(\nu_d^i\right) - \psi_1\left(\sum_{i=1}^{2} \nu_d^i\right)\right]$$

$$= \left(\delta_i + \sum_{t=1}^{T} \psi_{a_t}^i - \nu_d^i\right)\left[\psi_1\left(\nu_d^i\right) - \psi_1\left(\sum_{i=1}^{2} \nu_d^i\right)\right].$$

Set $\frac{dELBO}{d\nu_d^i} = 0$, as $\left[\psi_1\left(\nu_d^i\right) - \psi_1\left(\sum_{i=1}^{2} \nu_d^i\right)\right] > 0$ (Note $\psi' > 0, \psi'' < 0$), we obtain:

$$\nu_d^i = \delta_i + \sum_{t=1}^{T} \psi_{a_t}^i.$$

(11) Maximize with respect to $\boldsymbol{\tau_q}$

$$\frac{dELBO}{d\tau_q^{k,v}} = \left(\phi_q^{v,k} * num_q^v\right)\left[\psi_1\left(\tau_q^{k,v}\right) - \psi_1\left(\sum_{v=1}^{V} \tau_q^{k,v}\right)\right] + \left(\alpha_q^v - 1\right)\left[\psi_1\left(\tau_q^{k,v}\right) - \psi_1\left(\sum_{v=1}^{V} \tau_q^{k,v}\right)\right]$$

$$+ \frac{1}{B\left(\boldsymbol{\tau_q^k}\right)} B\left(\boldsymbol{\tau_q^k}\right)\left[\psi\left(\tau_q^{k,v}\right) - \psi\left(\sum_{v=1}^{V} \tau_q^{k,v}\right)\right] - \left[\psi\left(\tau_q^{k,v}\right) - \psi\left(\sum_{v=1}^{V} \tau_q^{k,v}\right)\right]$$

$$- \left(\tau_q^{k,v} - 1\right)\left[\psi_1\left(\tau_q^{k,v}\right) - \psi_1\left(\sum_{v=1}^{V} \tau_q^{k,v}\right)\right]$$

$$= \left(\phi_q^{v,k} * num_q^v + \alpha_q^v - \tau_q^{k,v}\right)\left[\psi_1\left(\tau_q^{k,v}\right) - \psi_1\left(\sum_{v=1}^{V} \tau_q^{k,v}\right)\right],$$

where num_q^v is the number of word v in document q.

Set $\frac{dELBO}{d\tau_q^{k,v}} = 0$, as $\left[\psi_1\left(\tau_q^{k,v}\right) - \psi_1\left(\sum_{v=1}^{V} \tau_q^{k,v}\right)\right] > 0$ (Note $\psi' > 0, \psi'' < 0$), we obtain $\tau_q^{k,v} = \phi_q^{v,k} * num_q^v + \alpha_q^v$. (Note that $\frac{d^2ELBO}{d\tau_q^{k,v^2}} = \left(\phi_q^{v,k} * num_q^v + \alpha_q^v - \tau_q^{k,v}\right)\left[\psi_2\left(\tau_q^{k,v}\right) - \psi_2\left(\sum_{v=1}^{V} \tau_q^{k,v}\right)\right] - \left[\psi_1\left(\tau_q^{k,v}\right) - \psi_1\left(\sum_{v=1}^{V} \tau_q^{k,v}\right)\right] = -\left[\psi_1\left(\tau_q^{k,v}\right) - \psi_1\left(\sum_{v=1}^{V} \tau_q^{k,v}\right)\right] < 0$.)

(12) Maximize with respect to $\boldsymbol{\tau_a}$

$$\begin{aligned}\frac{dELBO}{d\tau_a^{k,v}} = \sum_{t=1}^{T}\left(\phi_{a_t}^{v,k} * num_{a_t}^v\right)\left[\psi_1\left(\tau_a^{k,v}\right) - \psi_1\left(\sum_{v=1}^{V} \tau_a^{k,v}\right)\right] + (\alpha_a^v \\ - 1)\left[\psi_1\left(\tau_a^{k,v}\right) - \psi_1\left(\sum_{v=1}^{V} \tau_a^{k,v}\right)\right] + \frac{1}{B(\boldsymbol{\tau_a^k})} B\left(\boldsymbol{\tau_a^k}\right)\left[\psi\left(\tau_a^{k,v}\right) - \psi\left(\sum_{v=1}^{V} \tau_a^{k,v}\right)\right] \\ - \left[\psi\left(\tau_a^{k,v}\right) - \psi\left(\sum_{v=1}^{V} \tau_a^{k,v}\right)\right] - \left(\tau_a^{k,v} - 1\right)\left[\psi_1\left(\tau_a^{k,v}\right) - \psi_1\left(\sum_{v=1}^{V} \tau_a^{k,v}\right)\right] \\ = \left(\sum_{t=1}^{T} \phi_{a_t}^{v,k} * num_{a_t}^v + \alpha_a^v - \tau_a^{k,v}\right)\left[\psi_1\left(\tau_a^{k,v}\right) - \psi_1\left(\sum_{v=1}^{V} \tau_a^{k,v}\right)\right],\end{aligned}$$

where $num_{a_t}^v$ is the number of word v in document a_t.

Set $\frac{dELBO}{d\tau_a^{k,v}} = 0$, we obtain $\tau_a^{k,v} = \sum_{t=1}^{T} \phi_{a_t}^{v,k} * num_{a_t}^v + \alpha_a^v$.

If there are multiple Q&A documents, the updating formulas of $\boldsymbol{\tau_q}$ and $\boldsymbol{\tau_a}$ are as follows:

$$\tau_q^{k,v} = \sum_{d=1}^{D} \phi_{d,q}^{v,k} * num_{d,q}^v + \alpha_q^v,$$

$$\tau_a^{k,v} = \sum_{d=1}^{D}\sum_{t=1}^{T} \phi_{d,a_t}^{v,k} * num_{d,a_t}^v + \alpha_a^v.$$

B.3. A Variant of TM-OKC (one $\boldsymbol{\beta}$)

As we illustrate in Section 3.2 and Figure 1 of the main paper, our topic modeling framework TM-OKC allows questions and answers to have different topic-word distributions $\boldsymbol{\beta_q}$ and $\boldsymbol{\beta_a}$.

This is because the same topic in questions and answers can be expressed by different words. For example, in online news and comments, news is written by reporters while comments are made by the general public. For the same topic, words used in news can be more formal than those in comments. Therefore, we use different $\boldsymbol{\beta_q}$ and $\boldsymbol{\beta_a}$ to make the main framework as generalizable as possible. This setting is also seen in prior research (Ji et al. 2012). Note that if the texts are not observed, $\boldsymbol{\beta_q}$ and $\boldsymbol{\beta_a}$ are independent from each other. However, during the model inference, conditioned on the observed texts (i.e., $\boldsymbol{w_q}$ and $\boldsymbol{w_{a_t}}$), $\boldsymbol{\beta_q}$ and $\boldsymbol{\beta_a}$ are dependent due to the "*v-structures*" among parameters: "$\boldsymbol{z_q} \rightarrow \boldsymbol{w_q} \leftarrow \boldsymbol{\beta_q}$" and "$\boldsymbol{z_{a_t}} \rightarrow \boldsymbol{w_{a_t}} \leftarrow \boldsymbol{\beta_a}$" (Jordan 2003). Thus, $\boldsymbol{\beta_q}$ and $\boldsymbol{\beta_a}$ need to be jointly optimized to ensure their comparability. All the derivation of variational inference in Sections B.1 and B.2 of Web Appendix B is based on this general framework.

However, it should be noted that it makes more sense to adopt the same topic-word distribution for questions and answers in certain contexts (e.g., professional Q&A). Thus, we have intentionally created a variant with only one $\boldsymbol{\beta}$ (i.e., $\boldsymbol{\beta_q} = \boldsymbol{\beta_a}$). The graphical representation is shown in Figure B1.

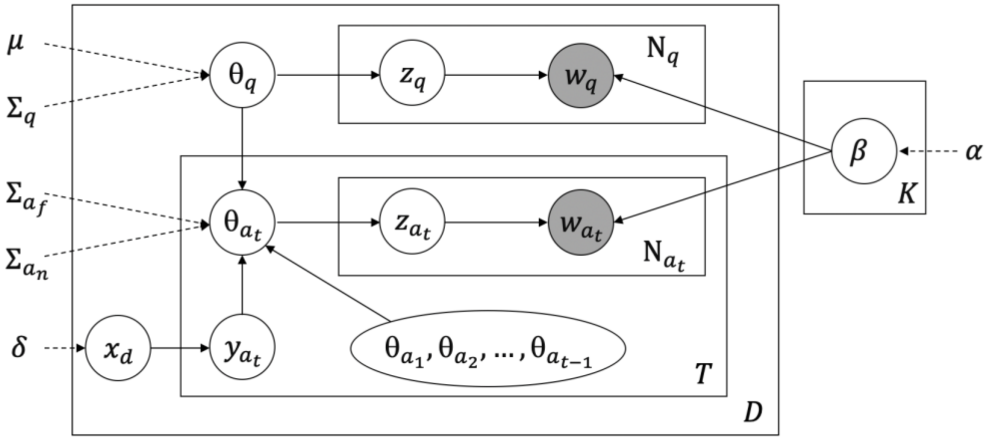


Figure B1: Graphical representation of the variant with one $\boldsymbol{\beta}$.

The model derivation can be obtained straightforwardly by revising the derivations in Sections B.1 and B.2 as follows.

(1) Substitute all $\boldsymbol{\beta_q}$ and $\boldsymbol{\beta_a}$ in the derivations with the same $\boldsymbol{\beta}$. And substitute the prior parameter of $\boldsymbol{\beta_q}$ and $\boldsymbol{\beta_a}$ (i.e., $\boldsymbol{\alpha_q}$ and $\boldsymbol{\alpha_a}$) with the same $\boldsymbol{\alpha}$.

(2) Substitute all $\boldsymbol{\tau_q}$ and $\boldsymbol{\tau_a}$ with the same $\boldsymbol{\tau}$. Recall that $\boldsymbol{\tau}$ is the parameter of the variational distribution of $\boldsymbol{\beta}$.

(3) Substitute the derivation of $\boldsymbol{\tau_q}$ and $\boldsymbol{\tau_a}$ in (11) and (12) in Section B.2 with the same derivation for $\boldsymbol{\tau}$ as follows.

$$
\begin{aligned}
\frac{dELBO}{d\tau_q^{k,v}} = {} & \left(\phi_q^{v,k} * num_q^v + \sum_{t=1}^{T} \phi_{a_t}^{v,k} * num_{a_t}^v\right)\left[\psi_1(\tau^{k,v}) - \psi_1\left(\sum_{v=1}^{V} \tau^{k,v}\right)\right] \\
& + (\alpha^v - 1)\left[\psi_1(\tau^{k,v}) - \psi_1\left(\sum_{v=1}^{V} \tau^{k,v}\right)\right] \\
& + \frac{1}{B(\boldsymbol{\tau}_{\boldsymbol{q}}^{\boldsymbol{k}})} B(\boldsymbol{\tau}_{\boldsymbol{q}}^{\boldsymbol{k}})\left[\psi(\tau^{k,v}) - \psi\left(\sum_{v=1}^{V} \tau^{k,v}\right)\right] - \left[\psi(\tau^{k,v}) - \psi\left(\sum_{v=1}^{V} \tau^{k,v}\right)\right] \\
& - (\tau^{k,v} - 1)\left[\psi_1(\tau^{k,v}) - \psi_1\left(\sum_{v=1}^{V} \tau^{k,v}\right)\right] \\
& = \left(\phi_q^{v,k} * num_q^v + \sum_{t=1}^{T} \phi_{a_t}^{v,k} * num_{a_t}^v + \alpha^v - \tau^{k,v}\right)\Bigg[\psi_1(\tau^{k,v}) \\
& - \psi_1\left(\sum_{v=1}^{V} \tau^{k,v}\right)\Bigg].
\end{aligned}
$$

where num_q^v is the number of word v in document q, and $num_{a_t}^v$ is the number of word v in document a_t.

Set $\frac{dELBO}{d\tau^{k,v}} = 0$, as $[\psi_1(\tau^{k,v}) - \psi_1(\sum_{v=1}^{V} \tau^{k,v})] > 0$ (Note $\psi' > 0, \psi'' < 0$), we obtain the updating formula of $\boldsymbol{\tau}$ as follows:

$$\tau^{k,v} = \phi_q^{v,k} * num_q^v + \sum_{t=1}^{T} \phi_{a_t}^{v,k} * num_{a_t}^v + \alpha^v.$$

If there are multiple Q&A documents, the updating formula of τ is as follows:

$$\tau^{k,v} = \sum_{d=1}^{D} \phi_{d,q}^{v,k} * num_{d,q}^v + \sum_{d=1}^{D} \sum_{t=1}^{T} \phi_{d,a_t}^{v,k} * num_{d,a_t}^v + \alpha^v.$$

B.4. Explanation of "Dependency and Variation"

Our framework is flexible in modeling the explicit structural relations by capturing both the dependency and variation within the Q&A thread. "Dependency" means that the topics of the current answer depend on the question and may also on prior answers, and "variation" means that users can focus on new topics that may not be saliently mentioned in the question or prior answers.

The mathematical details of the model structure are presented Section 3.2 of the main paper. Note that "dependent on" does not suggest "equal to" in the model. For example, for a novel answer, we draw the topic distribution $\boldsymbol{\theta}_{\boldsymbol{a_t}}$ from a *logistic-normal* distribution as follows:

$$\boldsymbol{\eta}_{\boldsymbol{a_t}} \sim N(\boldsymbol{\eta}_{\boldsymbol{q}}, \boldsymbol{\Sigma}_{\boldsymbol{a_n}}),$$

$$\boldsymbol{\theta}_{\boldsymbol{a_t}} = \frac{exp\ \{\boldsymbol{\eta}_{\boldsymbol{a_t}}\}}{\sum_{k=1}^{K} exp\ \{\eta_{a_t}^k\}}.$$

On the one hand, the mean parameter of this *logistic-normal* distribution is $\boldsymbol{\eta}_{\boldsymbol{q}}$ (the natural parameterization of the question's topic distribution $\boldsymbol{\theta}_q$), reflecting the dependency on the question. On the other hand, the variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{a_n}}$ reflects the variation. This is why the answer depends on the question while there are differences between the topic distributions of the question and its answer. It can be analogous to the familiar linear regression, $y = \beta \cdot x + \varepsilon,\ \varepsilon \sim N(0, \sigma)$. This regression is equivalent to $y \sim N(\beta \cdot x, \sigma)$, where y depends on $\beta \cdot x$ but is not equal to $\beta \cdot x$ because there is the variation parameter σ to capture the "fluctuation" around $\beta \cdot x$.

The rationale behind this "dependency and variation" structure is quite straightforward, because users usually read questions before providing their answers. Some answers may be highly correlated with the question, while others may be less or barely correlated with the question. Such a variation is captured by the $\boldsymbol{\Sigma}_{\boldsymbol{a_n}}$. Intuitively, we can think that the topic distribution of a novel answer varies or fluctuates around the topic distribution of its question.

References

Blei, D.M., and Lafferty, J.D. 2005. "Correlated Topic Models," in *Advances in Neural Information Processing Systems*, pp. 147–154.

Blei, D.M., and Lafferty, J.D. 2007. "A Correlated Topic Model of Science," *The Annals of Applied Statistics* (1:1), Institute of Mathematical Statistics, pp. 17–35.

Byrd, R.H., Lu, P., Nocedal, J., and Zhu, C. 1995. "A Limited Memory Algorithm for Bound Constrained Optimization," *SIAM Journal on Scientific Computing* (16:5), pp. 1190–1208.

Ji, Z., Xu, F., Wang, B., and He, B. 2012. "Question-Answer Topic Model for Question Retrieval in Community Question Answering," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pp. 2471–2474.

Jordan, M. I. (2003). "An Introduction to Probabilistic Graphical Models."

Roberts, M.E., Stewart, B.M., and Airoldi, E.M. 2016. "A Model of Text for Experimentation in the Social Sciences," *Journal of the American Statistical Association* (111:515), pp. 988–1003.

Wang, C., and Blei, D.M. 2013. "Variational Inference in Nonconjugate Models," *Journal of Machine Learning Research*, (14:1), pp. 1005–1031.

Wang, F., Zhang, J.L., Li, Y., Deng, K., and Liu, J.S. 2021. "Bayesian Text Classification and Summarization via A Class-Specified Topic Model," *Journal of Machine Learning Research* (22:89), pp. 1–48.

Web Appendix C: Hyper-Parameter Settings and Computational Resources

(1) Hyper-Parameter Settings

In the experiments of evaluating the statistical model fit, we set the common hyper-parameters to be the same for fair comparison. For other hyper-parameters that are specific to each method, we follow recommendations from the original papers, as well as performing a grid search to find relatively optimal settings. Specifically, the hyper-parameters of our TM-OKC and other topic models are set as follows.

- **Number of topics.** For all topic models, we adopt a widely used approach to select the optimal number of topics from a set of predefined numbers (i.e., 5, 10, 20, 40 and 80). That is, we train the model on the training set, choose the number of topics based on the log-likelihood on the validation set, and finally report the performance on the holdout test set (Griffiths and Steyvers 2004; Roberts et al. 2019; Bapna et al. 2019).
- **Stopping criteria.** All these methods optimize the model by iteratively improving the log-likelihood. We use the same stopping criteria in training, e.g., the log-likelihood between two consecutive iterations is less than a pre-defined threshold (i.e., 1e-5).
- **Prior distribution**. For the hyper-parameters of the prior *Dirichlet* and *Beta* distributions (i.e., α and δ in our model) in the topic models, we follow prior literature and set them to 0.1 (Griffiths and Steyvers 2004). We also perform a grid search over the range of [0.01, 1] and the results are similar.
- **Other hyper-parameters specific to some methods.** For the topic models combined with deep language models (i.e., NTM and SCHOLAR), we perform a grid search for the dimension of hidden embedding (the original papers of NTM and SCHOLAR both recommended 300) over the set of [50, 150, 300, 450, 600], for the learning rate during training (the original paper of NTM and SCHOLAR recommended 0.001 and 0.002, respectively) over the set of [0.01, 0.002, 0.001,

0.0005, 0.0001], and for the batch size during training (the original paper of NTM and SCHOLAR recommended 512 and 200, respectively) over the set of [32, 128, 200, 512, 1024]. For LeadLDA, since it uses Gibbs sampling for model inference, we perform a grid search over the set of [500, 1000, 2000] for the maximum number of iterations and the results are similar.

In the experiments of *document classification* in the additional evaluation presented in Web Appendix F, we need to specify the hyper-parameters for the random forest algorithm where we feed the document-level topic vectors into a random forest model to predict the category of each document. Specifically, we perform a grid search over the set of [50, 100, 200, 300, 500] for the number of estimators, and over the set of [2, 4, 6, 8, 12, None] for the maximum depth of the decision trees. In addition, we follow the default settings of the scikit-learn package in Python for other hyper-parameters of the random forest algorithm.

(2) Computational Resources

All the experiments of are conducted on a machine with an Intel 8-core i9 CPU with 64GB of RAM. In our Stack Exchange datasets, the largest one is the category of “English Language & Usage” with 109,977 questions and 268,356 answers, which took approximately 23 hours to finish training our TM-OKC; the smallest one is the category of “Project Management” with 5,782 questions and 17,724 answers, which took approximately 0.7 hours to finish training our TM-OKC. Note that since we use EM algorithm, the most time-consuming part is the E-step, which can be parallelized with multiple processes to further reduce the run time.

References

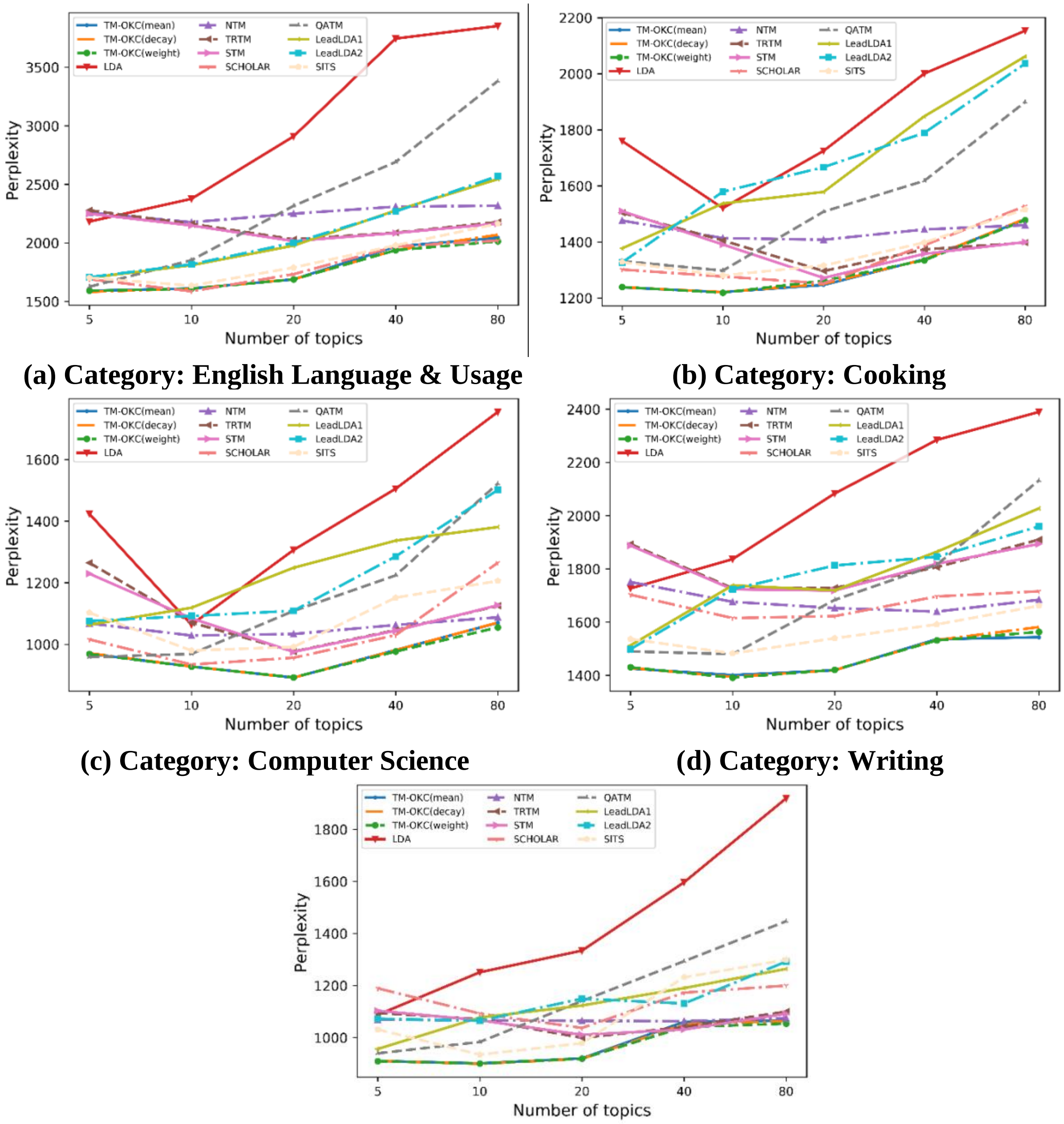
Bapna, S., Benner, M.J., and Qiu, L. 2019. “Nurturing Online Communities: An Empirical Investigation,” *MIS Quarterly* (43:2), pp. 425–452.

Griffiths, T.L., and Steyvers, M. 2004. “Finding Scientific Topics,” *Proceedings of the National Academy of Sciences*, (101: suppl_1), pp. 5228–5235.

Roberts, M.E., Stewart, B.M., and Tingley, D. 2019. “Stm: An R Package for Structural Topic Models,” *Journal of Statistical Software* (91:1), pp. 1–40.

Web Appendix D: Perplexity Scores for Other Categories

To show robustness of the model fit, we pictorially show the perplexity scores for each method (ours and the baselines) under different numbers of topics (i.e., 5, 10, 20, 40 and 80) on the holdout data.



(a) Category: English Language & Usage

(b) Category: Cooking

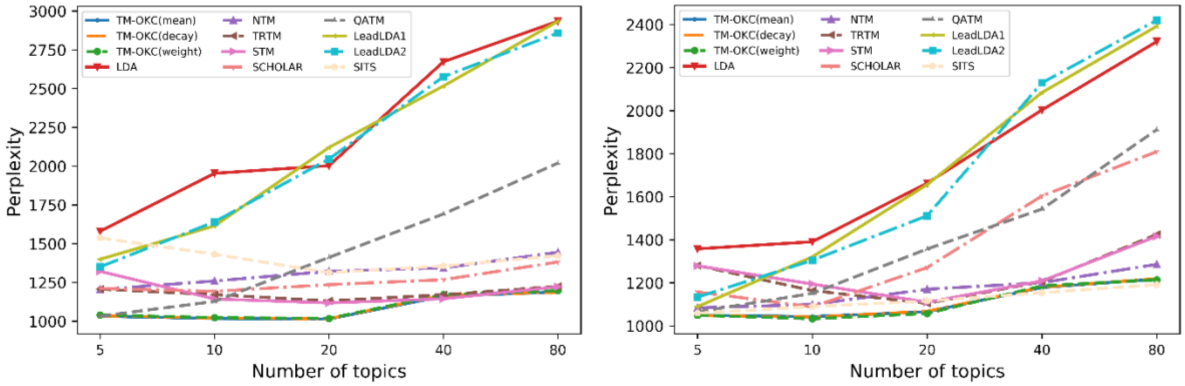
(c) Category: Computer Science

(d) Category: Writing

(e) Category: Project Management

Figure D1: Perplexity for other categories within Stack Exchange.

Figure D1 shows the comparison of different models for categories within Stack Exchange other than Data Science, which is shown in the main paper. Figure D2 shows the comparison of different models for categories within Quora other than Business and Marketing.



(a) Category: Science and Technology **(b) Category: Health and Life**

Figure D2: Perplexity for other categories within Quora.

Web Appendix E: Examination of Important Parameters in the TM-OKC

The TM-OKC models the interdependency of a question and its answers as well as the temporal variation of threaded answers in its Bayesian framework, which is significantly different from prior studies. For example, the TM-OKC allows differential impacts of the question and of prior answers on the current answer by incorporating the parameter γ. The TM-OKC also captures the variances of topic distributions for questions and answers, denoted by $\boldsymbol{\Sigma}_q$, $\boldsymbol{\Sigma}_{\boldsymbol{a_f}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{a_n}}$, respectively. Larger γ indicates that answers are more easily affected by prior answers. Larger $\boldsymbol{\Sigma}_q$ indicates more variation in the topics of questions, while larger $\boldsymbol{\Sigma}_{\boldsymbol{a_f}}$ or $\boldsymbol{\Sigma}_{\boldsymbol{a_n}}$ means that the topic distribution of follow-up or novel answers is more likely to fluctuate. These unique parameters in our model can reflect important structural information of OKC texts, which can potentially be used in subsequent empirical studies to generate new insights.

Table E1 shows values of γ learned from the Stack Exchange dataset, upon which we make the following observations. First, the values of γ are larger for the “soft skill” categories (i.e., Project Management, Writing, English Language & Usage and Cooking) compared to the “hard skill” categories (i.e., Data Science, Computer Science), suggesting that answers in soft skill categories are more likely to be affected by previous answers. This is expected because answers to questions in the soft skill categories are more flexible. Second, the values of γ for the categories of Data Science and Computer Science are less than 1 while the values for the other four categories are greater than 1, which indicates that topics of follow-up answers in these four categories might be dominated by the previous answers but not the original question. Table E2 shows values of γ learned from the Quora dataset, which are much higher than the values learned from the Stack Exchange dataset in general. This might be because Stack Exchange is a professional Q&A site while Quora is more like a social media platform, so the users of Stack

Exchange will focus more on answering the questions rather than on joining previous discussions.

Table E1: The learned γ under different models across categories of Stack Exchange.

Model	**Technology (Data Science)**	**Culture/ Recreation (English Language & Usage)**	**Life/Arts (Cooking)**	**Science (Computer Science)**	**Professional (Writing)**	**Business (Project Management)**
TM-OKC (mean)	0.38	2.72	2.59	0.89	2.68	2.65
TM-OKC (decay)	0.36	2.45	2.58	0.87	9.76	2.51
TM-OKC (weight)	0.72	2.66	1.90	0.77	2.04	2.08

Table E2: The learned γ under different models across categories of Quora.

Model	**Science and Technology**	**Business and Marketing**	**Health and Life**
TM-OKC (mean)	2.50	3.59	3.84
TM-OKC (decay)	2.31	3.75	3.98
TM-OKC (weight)	2.20	3.81	3.73

Table E3 summarizes the variances reflected by $\mathbf{\Sigma}_{\boldsymbol{a}_f}$ and $\mathbf{\Sigma}_{\boldsymbol{a}_n}$[1] learned from the Stack Exchange dataset, upon which we can make several notable observations. First, in the categories of Data Science and Computer Science, γ is smaller but $\mathbf{\Sigma}_{\boldsymbol{a}_f}$ and $\mathbf{\Sigma}_{\boldsymbol{a}_n}$ are larger. One possible explanation is that, in these two hard skill categories, users need to raise distinct topics to address hardcore technical issues. Therefore, users are more likely to adjust the focus of the discussion, driven by their intention to provide professional answers rather than participating for fun.

[1] We calculate the trace of the variance-covariance matrix divided by the dimension in order to measure the average variance of topic distribution per dimension.

Second, in the two hard skill categories, $\boldsymbol{\Sigma}_{a_f}$ is larger than $\boldsymbol{\Sigma}_{a_n}$, indicating greater fluctuation in follow-up answers than novel answers. This might be because a novel answer tends to address the technical question directly and thus would not deviate much, while a follow-up answer may deviate from the previous discussions due to the user's own interests. However, in the categories of English Language & Usage and Cooking, $\boldsymbol{\Sigma}_{a_f}$ can be smaller than $\boldsymbol{\Sigma}_{a_n}$ because the follow-up discussions in these categories are more like daily chat and thus do not fluctuate much. In addition, Table E4 summarizes the variance parameters learned from the Quora dataset. We can see that the heterogeneity among different Quora categories is not very significant compared to Stack Exchange, which shows the different styles of these two Q&A platforms.

Table E3: The variances of topic distribution by different models across Stack Exchange categories.

	Model	**Technology (Data Science)**	**Culture/ Recreation (English Language & Usage)**	**Life/Arts (Cooking)**	**Science (Computer Science)**	**Professional (Writing)**	**Business (Project Management)**
$\boldsymbol{\Sigma}_{a_f}$	TM-OKC (mean)	5.96	0.46	0.48	4.12	1.72	0.49
	TM-OKC (decay)	3.86	0.49	0.48	4.51	1.63	0.49
	TM-OKC (weight)	3.52	0.52	0.53	7.33	0.50	0.51
$\boldsymbol{\Sigma}_{a_n}$	TM-OKC (mean)	1.17	0.74	0.82	1.27	0.59	0.73
	TM-OKC (decay)	2.32	0.78	0.81	1.27	0.60	0.74
	TM-OKC (weight)	1.14	0.74	0.82	1.26	0.71	0.73

Table E4: The variances of topic distribution by different models across Quora categories.

	Model	Science and Technology	Business and Marketing	Health and Life
$\boldsymbol{\Sigma}_{\boldsymbol{a_f}}$	TM-OKC (mean)	0.60	0.61	0.57
	TM-OKC (decay)	0.61	0.59	0.56
	TM-OKC (weight)	0.63	0.57	0.59
$\boldsymbol{\Sigma}_{\boldsymbol{a_n}}$	TM-OKC (mean)	0.98	1.08	1.04
	TM-OKC (decay)	1.01	1.06	1.01
	TM-OKC (weight)	1.03	1.05	1.06

Web Appendix F: Additional Evaluation of the TM-OKC

In this Appendix, we present the details about the additional evaluation of our TM-OKC in terms of representation capability and interpretability.

F.1. Representation Capability

We use a prediction task of *document classification* to demonstrate the representation capability among different methods; this has been used by previous studies to evaluate the effectiveness of topic models (Zeng et al. 2019; Yang et al. 2022). Specifically, the prediction task is formalized in three steps. (1) Sample 2,000 questions and their answers from each of the six categories in our Stack Exchange dataset (or sample 1,000 questions and their answers from each of the three categories within our Quora dataset), then combine them to form a dataset for document classification, where the category is the label in this supervised classification task. (2) Apply topic models on the combined dataset to obtain a topic vector for each document. (3) Feed the learned topic vector to a classifier (i.e., random forest) to predict the document category.

Baselines: To validate the effectiveness of our method, we choose three sets of baselines to obtain document representations (i.e., step 2 as described above).

- Topic models. We use all benchmark topic models listed in Section 5 (i.e., LDA, NTM, TRTM, STM, SCHOLAR, QATM, LeadLDA and SITS).
- Basic textual feature extraction methods. We employ the commonly used *term frequency–inverse document frequency* (TF-IDF) method with the top *W* (*W*=100, 500 or 1,000) words in the corpus to represent each document.
- Representation learning methods. First, we choose bi-directional long-short term memory (Bi-LSTM), given its success in many document classification tasks (Adhikari et al. 2019). Second, we choose the transformer-based BERT, an advanced large language model (Devlin

et al. 2019). To make a comprehensive comparison, we use both pre-trained and fine-tuned BERT[2]. Note that although they may have strong prediction power, the representation learning methods are not able to produce interpretable results (e.g., topics). While our study focuses on unsupervised topic modeling that can extract interpretable topics from texts, we still include these cutting-edge representation learning methods for comparison to show the representative capability of TM-OKC.

Evaluation Metrics: Since this is a standard multi-class classification task and the datasets are balanced, we use the standard classification accuracy (i.e., the percentage of correctly classified instances) as the evaluation metric. We repeat the experiments 30 times and report the average performance.

The prediction results are presented in Table F1. For topic modeling methods, we vary the number of topics to show robustness. From the table, we make the following observations. First, our model achieves the best prediction performance among all topic modeling methods across different numbers of topics. Second, all topic modeling methods other than LDA show better performance than the basic TF-IDF feature extraction method. Third, using the topic vectors derived from our topic model as the input to a simple machine learning classifier (i.e., random forest) can achieve even better performance than Bi-LSTM and comparable performance with the fine-tuned BERT. These results highlight the fact that topic models can be used to represent semantics of texts, and compared to existing topic modeling methods, our model has stronger representation capability on OKC texts.

[2] For pre-trained BERT, we obtain the document vector (e.g., associated with the CLS token) as input and feed it to a classifier (i.e., random forest) to predict the document label; for Bi-LSTM and fine-tuned BERT, models are tuned using labeled supervision under the document classification task.

Table F1: Prediction accuracy of different methods on two modified datasets.

		Stack Exchange			Quora		
	Number of topics	40	80	120	40	80	120
Basic text feature extraction methods	TF-IDF features (top 100 words)		0.517			0.600	
	TF-IDF features (top 500 words)		0.717			0.779	
	TF-IDF features (top 1000 words)		0.718			0.804	
Bayesian topic modeling methods	LDA	0.727	0.710	0.699	0.728	0.726	0.712
	TRTM	0.891	0.885	0.893	0.935	0.920	0.937
	STM	0.886	0.847	0.888	0.935	0.935	0.930
	QATM	0.898	0.853	0.860	0.960	0.962	0.937
	LeadLDA1	0.867	0.832	0.751	0.762	0.749	0.747
	LeadLDA2	0.865	0.865	0.749	0.796	0.762	0.737
	SITS	0.903	0.891	0.868	0.881	0.893	0.876
Topic modeling combined with deep language models	NTM	0.905	0.900	0.859	0.813	0.827	0.848
	SCHOLAR	0.901	0.903	0.885	0.826	0.855	0.818
Representation learning methods	Bi-LSTM		0.822			0.919	
	Pre-trained BERT		0.696			0.722	
	Fine-tuned BERT		0.936			0.992	
Our method	TM-OKC	0.911*	0.933***	0.918***	0.983**	0.990***	0.968***

Note that the statistical significance is calculated compared with the best topic modeling method under a one-tailed *t*-test. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

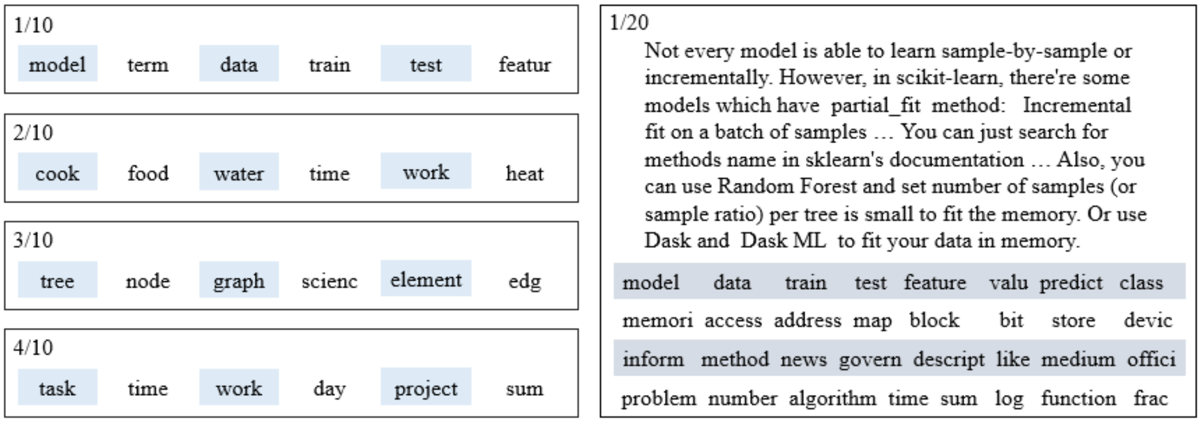
F.2. Interpretability

We now turn to investigating the interpretability of the generated topics. One intuitive way to do this is by evaluating face validity, where we show the top 10 words for several topics of our model and the best baseline model (please refer to Web Appendix G). To further explore the interpretability in a more rigorous way, we follow previous research (Chang et al. 2009; Bao and Datta 2014; Palese and Piccoli 2020) and conduct *word intrusion* and *topic intrusion* through lab studies. The detailed procedures are described below.

The *word intrusion* task is used to quantitatively measure the semantic coherence of the identified topics. Specifically, the subject is presented with six randomly ordered words. The task of the subject is to find the one word (i.e., the *intruder*) that is out of place or not in line with the others. If the set of words excluding the intruder makes sense together, then the subject can easily identify the intruder. For example, in the set {elephant, tiger, horse, apple, pig, cow}, most people can identify “apple” as the *intruder* because the remaining words, {elephant, tiger, horse, pig, cow} are coherent – they are all animals. In contrast, for another set {car, teacher, lion, agile, blue, square} that lacks such semantic coherence, it is difficult to identify the *intruder*. In order to evaluate the semantic coherence of a topic, we first select the five most probable words of a topic. Then, an intruder word is randomly selected from those words with low probabilities in the current topic (to reduce the possibility of the *intruder* coming from the same semantic group) but high probabilities in some other topics (to ensure that the *intruder* would not be rejected solely because of rarity). Finally, all six words are shuffled and presented to the subject like in Figure F1-(a). The evaluation metric for such a *word intrusion* task is the model precision, which is defined as the fraction of subjects agreeing with the topic model. Specifically, the word intrusion precision of the *k*-th topic learned by model *m* is defined as:

$$WIP_m^k = \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}(i_{k,s}^m = w_k^m),$$

where w_k^m is the true intruding word among the set of words generated from the *k*-th topic learned by model *m*, $i_{k,s}^m$ is the intruder selected by subject *s* from the set of words generated from the *k*-th topic learned by model *m*, *S* is the number of subjects, and $\mathbf{1}(\cdot)$ is the indicator function. Finally, the precision of model *m* (i.e., WIP_m) is obtained by taking the average of WIP_m^k over topics.



1/10

model term data train test featur

2/10

cook food water time work heat

3/10

tree node graph scienc element edg

4/10

task time work day project sum

1/20

Not every model is able to learn sample-by-sample or incrementally. However, in scikit-learn, there're some models which have partial_fit method: Incremental fit on a batch of samples … You can just search for methods name in sklearn's documentation … Also, you can use Random Forest and set number of samples (or sample ratio) per tree is small to fit the memory. Or use Dask and Dask ML to fit your data in memory.

model data train test feature valu predict class

memori access address map block bit store devic

inform method news govern descript like medium offici

problem number algorithm time sum log function frac

(a) *Word intrusion* **(b) *Topic intrusion***

Figure F1: Screenshots of the lab studies for *word intrusion* and *topic intrusion*.

The *topic intrusion* task measures whether the model's document decomposition, in which a document is broken down into a mixture of topics, is consistent with human judgment of the same document. In this task, human subjects are presented with a document along with four topics (with each topic represented by the eight words with highest probabilities within that topic), as in Figure F1-(b). Three of those topics are those with the highest probabilities associated with that document. The remaining intruder topic is randomly chosen from other topics with low probabilities. The evaluation metric for *topic intrusion* task is *topic log odds*, a quantitative measure of the agreement between the model and human judgment. Specifically, the measure is defined as the log ratio of the probability assigned to the true intruder to the probability assigned to the intruder selected by the subject:

$$TLO_m^d = \frac{1}{S}\sum_{s=1}^{S}\left(\log\hat{\theta}^m_{d,t_d^m} - \log\hat{\theta}^m_{d,j^m_{d,s}}\right),$$

where t_d^m is the true intruder topic among those with highest probabilities of document *d* inferred by model *m*, $j^m_{d,s}$ is the intruder topic selected by subject *s* from the topics with highest probabilities of document *d* inferred by model *m*, $\hat{\theta}^m_{d,k}$ is the probability assigned to topic *k* in document *d* inferred by model *m*, and *S* is the number of subjects. Finally, the overall measure for model m (i.e., TLO_m) is obtained by taking the average of TLO_m^d over documents. A larger value of TLO_m indicates a greater agreement between the judgment of the model and the subjects. We can see that the upper bound of TLO_m is 0, which can only be achieved when all subjects pick out the true intruder topics for all documents.

Table F2: Human evaluation results of *word intrusion* and *topic intrusion* tasks.

	WIP_m in *word intrusion*			TLO_m in *topic intrusion*		
Number of topics	40	80	120	40	80	120
LDA	0.806	0.794	0.788	-1.61	-1.65	-1.85
NTM	0.825	0.813	0.819	-1.52	-1.47	-1.48
TRTM	0.813	0.838	0.806	-1.43	-1.35	-1.49
STM	0.819	0.825	0.819	-1.37	-1.25	-1.47
SCHOLAR	0.813	0.819	0.813	-1.36	-1.30	-1.51
QATM	0.825	0.831	0.813	-1.31	-1.22	-1.39
SITS	0.819	0.825	0.800	-1.32	-1.29	-1.38
LeadLDA1	0.813	0.806	0.800	-	-	-
LeadLDA2	0.819	0.813	0.794	-	-	-
TM-OKC	0.856^{**}	0.869^{**}	0.838^{*}	-1.12^{**}	-0.96^{***}	-1.23^{**}

Note: (1) The statistical significance is calculated compared with the best topic modeling method under a one-tailed *t*-test. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; (2) As LeadLDA only assigns one single topic to each post, the TLO_m cannot be calculated and thus LeadLDA is excluded in the *topic intrusion* task.

Using the same two modified datasets introduced in Section F.1, we conduct the two tasks separately on our model and the baseline topic models (i.e., LDA, NTM, TRTM, STM, SCHOLAR, QATM, LeadLDA and SITS). For the *word intrusion* task, we evaluate the top 10 topics generated by each topic model on two datasets across different numbers of topics (i.e., 40, 80 and 120). For the *topic intrusion* task, we randomly sample 60 posts from the corpus and evaluate the performance for each topic model across different numbers of topics. To carry out these tasks with human subjects, we use the popular crowdsourcing platform Amazon Mechanical Turk and present each subject with 10 *word intrusion* or 20 *topic intrusion* tasks. For the sake of robustness, we ensure each task is performed by eight different workers (Chang et al.

2009). The results are shown in Table F2. From the table, we find that our model achieves significantly better human evaluation performance.

References

Adhikari, A., Ram, A., Tang, R., and Lin, J. 2019. “Rethinking Complex Neural Network Architectures for Document Classification,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (1), pp. 4046–4051.

Bao, Y., and Datta, A. 2014. “Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures,” *Management Science* (60:6), pp. 1371–1391.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., and Blei, D.M. 2009. “Reading Tea Leaves: How Humans Interpret Topic Models,” in *Advances in Neural Information Processing Systems*.

Devlin, J, Chang, M.W., Lee, K., and Toutanova K. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding,” in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Tech. (Association for Computational Linguistics, Stroudsburg, PA)*, pp. 4171–4186.

Palese, B., and Piccoli, G. 2020. “Evaluating Topic Modeling Interpretability Using Topic Labeled Gold-Standard Sets,” *Communications of the Association for Information Systems* (47), pp. 433–451.

Yang, Y., Zhang, K., and Fan, Y. 2022. “sDTM: A Supervised Bayesian Deep Topic Model for Text Analytics,” *Information Systems Research*.

Zeng, J., Li, J., He, Y., Gao, C., Lyu, M.R., and King, I. 2019. “What You Say and How You Say It: Joint Modeling of Topics and Discourse in Microblog Conversations,” *Transactions of the Association for Computational Linguistics* (7), pp. 267–281.

Web Appendix G: Face Validity of Generated Topics

To provide the face validity of generated topics, we show the top 10 words with highest probabilities in the top 10 topics of our model and of the best baseline model.

For the modified Stack Exchange dataset, we choose NTM as the baseline, and we report the results of our model and NTM with 40 topics in Table G1 and Table G2. This is because NTM achieves the best prediction performance among the topic model baselines in the *document classification* task in Section F.1, with 40 topics. As shown in Table G1 and G2, Topic 7 and Topic 8 learned by NTM are both about "model" and are not very distinguishable, while the topic about "model" is well captured in one single topic (i.e., Topic 4) of TM-OKC.

Table G1: Top 10 topics learned by TM-OKC on the Stack Exchange dataset.

Topic	Most probable words
1	quot, charact, stori, write, reader, word, think, thing, way, peopl
2	cook, food, water, time, heat, temperatur, pan, oil, get, meat
3	team, product, work, sprint, scrum, stori, develop, agil, user, backlog
4	model, data, train, test, featur, valu, predict, class, learn, set
5	project, manag, need, work, risk, cost, time, peopl, develop, requir
6	recip, flour, bake, dough, egg, sugar, bread, milk, tast, flavor
7	problem, number, algorithm, time, sum, log, function, frac, comput, solv
8	languag, state, word, quot, type, machin, context, mean, accept, definit
9	tree, node, graph, element, edg, array, vertex, algorithm, path, number
10	task, time, work, day, project, hour, resourc, date, start, schedul

Table G2: Top 10 topics learned by NTM on the Stack Exchange dataset.

Topic	Most probable words
1	team, project, work, scrum, sprint, time, task, need, product, process
2	cook, water, add, time, food, pan, oil, flour, good, need
3	character, reader, write, think, want, know, thing, way, need, good
4	quot, word, english, mean, phrase, verb, say, noun, think, know
5	problem, number, algorithm, time, set, sum, give, function, log, find
6	write, book, work, good, read, want, find, publish, page, author
7	model, time, algorithm, problem, need, good, way, number, case, set
8	model, train, test, class, quot, dataset, layer, input, set, loss
9	work, time, need, know, way, good, want, find, question, write
10	time, work, need, know, way, good, want, find, question, quot

For the Quora dataset, we report the results of our model and QATM with 80 topics in Table G3 and Table G4. This is because QATM achieves the best prediction performance among the topic model baselines in the *document classification* task of Section F.1, with 80 topics. It can be seen that the topic about "code and software" spreads in three topics (Topic 2, 5 and 10) in QATM, while this topic is well captured in one single topic (i.e., Topic 3) in TM-OKC.

Table G3: Top 10 topics learned by TM-OKC on the Quora dataset.

Topic	Most probable words
1	time, life, peopl, get, think, good, know, thing, even, day
2	bitcoin, invest, market, cryptocurr, crypto, buy, coin, trade, money, ethereum
3	softwar, engin, work, code, develop, program, test, problem, need, job
4	peopl, thing, question, find, need, ask, want, know, answer, think
5	wear, look, woman, cloth, dress, shirt, style, girl, jean, fashion
6	cost, pay, countri, govern, increas, rate, inflat, high, economi, tax
7	god, human, differ, truth, philosophi, religion, world, analysi, exist, purpos
8	content, medium, websit, seo, blog, social, search, googl, page, post
9	busi, product, brand, custom, start, onlin, digit, servic, company, plan
10	weight, eat, lose, exercis, food, calori, diet, bodi, healthi, day

Table G4: Top 10 topics learned by QATM on the Quora dataset.

Topic	Most probable words
1	bitcoin, invest, cryptocurr, crypto, ethereum, buy, market, blockchain, transact, time
2	time, work, softwar, know, need, way, code, someth, develop, think
3	wear, look, dress, tri, get, cloth, string, time, good, love
4	god, energi, human, object, exist, say, time, peopl, medit, self
5	engin, code, program, comput, softwar, problem, languag, time, write, thing
6	develop, applic, busi, compani, get, work, start, need, peopl, idea
7	unit, peopl, implement, test, chang, mani, function, may, thing, engin
8	project, request, screen, video, get, creat, need, want, experi, manag
9	work, system, way, program, task, time, need, weight, get, lose
10	code, test, softwar, interview, engin, question, skill, manag, work, ask

From these tables, we can see that in comparison to the best benchmark, the topics learned by our TM-OKC model are more coherent and distinguishable, providing a face validity of its effectiveness in modeling interdependencies among questions and answers in OKCs.

Web Appendix H: Logic and Additional Evaluation of User Profiling

(1) Logic of the User Profiling Example

As illustrated in Figure 3 of the main paper, with *better statistical model fit*, *representation capability* and *interpretability*, our TM-OKC can benefit many downstream tasks, which can be user-related (e.g., user profiling) or not user-related (e.g., trending topic detection). In our study, user profiling is selected as an example to demonstrate the practical utility and relative merit of TM-OKC.

As elaborated in Section 2 of the main paper, our study aims to developing a general topic modeling framework that explicitly captures the complex structural relationships among OKC texts, thus we do not model the observed attributes of texts (e.g., authorship information) which is beyond our research focus. Note that although this user profiling example happens to be user-related, it does not mean we have to model authorship a priori. This is because the key of topic models is to obtain good text representation (i.e., topic vectors), and this modeling process does not necessarily include authorship information. After deriving topic vectors from texts, different downstream tasks can use these topic vectors in their own ways. For example, here in this study, we use topic vectors to construct user profiles. This procedure actually incorporates authorship information a posteriori. Other topic models that capture authorship a priori follow the same procedure. However, this is just a downstream task, which does not impose restrictions on whether the topic model should include authorship information a priori. This is also reflected in prior research. For example, LDA, which does not model authorship information, has also been applied to construct Twitter users' online profiles (Geva et al. 2019). In addition, although modeling authorship is beyond our research scope, we still compared our method with the state-of-the-art baseline methods that model authorship and achieved significantly better performance, through which we have empirically demonstrated the importance of modeling explicit structural relationships among OKC texts and made our methodological contributions in this regard. It is worth noting that adding authorship into our framework might be able to further improve the model performance, which we leave for future research.

(2) Additional Evaluation

In the user profiling experiments in Section 6 of the main paper, we follow the strategy used in prior studies to perform a similarity search in a 100-question pool randomly selected from the hold-out test set (Elkahky et al. 2015; He et al. 2017). Here we also present the results using the whole hold-out test set as the question pool to perform the similarity search in Tables H1 and H2.

Table H1: User profiling performance comparison of different methods under a moderate data size (number of Q&A threads is 8,000).

		Hit rate for top K		
		K=5	K=10	K=20
Basic text feature extraction methods	TF-IDF features (top 100 words)	0.17%	0.34%	0.63%
	TF-IDF features (top 500 words)	0.15%	0.26%	0.52%
	TF-IDF features (top 1000 words)	0.17%	0.27%	0.63%
Bayesian topic modeling methods	LDA	0.83%	1.92%	3.95%
	TRTM	0.95%	2.11%	4.03%
	STM	1.02%	2.30%	4.23%
	QATM	1.07%	2.19%	4.11%
	LeadLDA1	0.17%	0.31%	0.63%
	LeadLDA2	0.19%	0.37%	0.78%
	SITS	1.16%	2.29%	4.23%
Pre-trained deep language models	Pre-trained BERT	0.48%	0.84%	1.45%
Topic modeling combined with deep language models	NTM	0.79%	1.90%	3.39%
	SCHOLAR	0.91%	1.98%	4.08%
Neural matrix factorization	NMF	0.88%	1.83%	3.90%
Our method	TM-OKC	1.28%**	2.55%**	4.58%**

Note that the statistical significance is calculated compared with the best baseline method under a one-tailed *t*-test. ** $p < 0.05$, *** $p < 0.01$.

Not surprisingly, the results presented in Tables H1 and H2 are consistent with the results presented in Tables 10 and 11 in terms of the relative performance across different methods[3].

[3] Note that in Table H2, the hit rate for top K (K=10) decreases with the increase of data size. This is because the size of the question pool (i.e., the whole test set in this case) used in similarity search increases with the data size, which makes it more difficult to find the true question answered by a specific user. However, the relative trend between the performance of our TM-OKC and that of other methods remains consistent with what we see in Table 11.

Table H2: User profiling performance comparison of different methods under different data sizes for *K*=10 (i.e., top 10 hit rate).

		Data size *N* (number of Q&A threads)						
		1,000	2,000	4,000	8,000	16,000	32,000	64,000
Basic text feature extraction methods	TF-IDF features (top 100 words)	1.19%	0.73%	0.60%	0.34%	0.27%	0.17%	0.12%
	TF-IDF features (top 500 words)	1.37%	0.80%	0.66%	0.26%	0.31%	0.20%	0.13%
	TF-IDF features (top 1000 words)	1.37%	0.87%	0.61%	0.27%	0.24%	0.15%	0.13%
Bayesian topic modeling methods	LDA	5.13%	3.74%	2.40%	1.92%	1.37%	0.74%	0.65%
	TRTM	5.96%	4.02%	2.54%	2.11%	1.41%	0.90%	0.72%
	STM	6.45%	4.12%	2.53%	2.30%	1.55%	0.96%	0.79%
	QATM	6.26%	4.30%	2.40%	2.19%	1.52%	1.05%	0.79%
	LeadLDA1	1.31%	0.72%	0.44%	0.31%	0.19%	0.24%	0.19%
	LeadLDA2	1.50%	0.75%	0.68%	0.37%	0.35%	0.25%	0.21%
	SITS	5.41%	4.09%	2.69%	2.29%	1.61%	1.08%	0.79%
Pre-trained deep language models	Pre-trained BERT	2.44%	1.87%	1.06%	0.84%	0.71%	0.61%	0.59%
Topic modeling combined with deep language models	NTM	2.32%	2.22%	1.78%	1.90%	1.74%	1.28%	1.04%
	SCHOLAR	4.82%	3.07%	2.10%	1.98%	1.77%	1.30%	1.10%
Neural matrix factorization	NMF	2.11%	1.97%	1.70%	1.83%	1.68%	1.23%	1.02%
Our method	TM-OKC	7.33%***	4.51%***	2.98%**	2.55%**	1.98%**	1.37%*	1.10%

Note that the statistical significance is calculated compared with the best baseline method under a one-tailed *t*-test. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

References

Elkahky, A. M., Song, Y., and He, X. 2015. "A Multi-View Deep Learning Approach for Cross Domain User Modeling in Recommendation Systems," in *Proceedings of the 24th International Conference on World Wide Web*, pp. 278–288.

Geva, H., Oestreicher-Singer, G., and Saar-Tsechansky, M. 2019. "Using Retweets When Shaping Our Online Persona: Topic Modeling Approach," *MIS Quarterly* (43:2), pp. 501–524.

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., and Chua, T. S. 2017. "Neural Collaborative Filtering," in *Proceedings of the 26th International Conference on World Wide Web*, pp. 173–182.